

THIS WEEK

EDITORIALS

PUBLISHING Nature journals to offer double-blind peer review **p.274**

WORLD VIEW Trouble ahead for the European Spallation Source **p.275**



REPEL BOARDERS Thirsty toads offer fresh strategy to combat invaders **p.277**

Beyond the genome

Studies of the epigenomic signatures of many healthy and diseased human tissues could provide crucial information to link genetic variation and disease.

The Greek prefix *epi-* can signify upon, on, over, near, at, before, and after. Most of those could apply to its use in the term ‘epigenetics’ — particularly the last of them. It is some 14 years, almost to the day, that *Nature* published the draft sequence of the human genome. Now, in this issue, we publish results from a subsequent study on the non-genetic modifications to the genome — epigenetic modifications — that crucially determine which genes are expressed by which cell type, and when.

It is hard to think of any branch of human biology that has not benefited from the human genome sequence. Its legacy has perhaps been most notable in advances in our appreciation of the part that genetics and genetic variation play in the normal functioning of a human body and in disease. But despite the progress, each question that the genome helps to answer throws up further questions. Much remains to be understood about how genetic information is interpreted by the individual cells in our body.

This is where epigenetics comes in. Upon the genome, on the genome, over the genome — take your pick — epigenetics collectively describes changes in the regulation of gene expression that can be passed on to a cell's progeny but are not due to changes to the nucleotide sequence of the gene.

Soon after the human genome sequence had been completed, it became clear that an epigenome — a map of the genome-wide modifications made to DNA and the protein scaffold that supports it — would also be required. The task at hand was, as researchers like to say, not trivial. Every cell in the body carries the same genome (with a few exceptions), but the epigenome changes with cell and tissue type.

Epigenetics is still an emerging science, but researchers are now building tools to study epigenetic changes in the genome in a systematic and genome-wide way. In 2012, *Nature* celebrated the publication of the results of the ENCODE project, the aim of which was to describe all the functional elements encoded in the human genome by mapping epigenetic modifications (see nature.com/encode).

ENCODE was a pioneer in scale of effort and development of specialized analytical software, and has already had a tremendous impact on human-genetics studies. But its clinical application is limited because most of its results come from a small number of laboratory cell lines. Clinically useful epigenetic information must instead be drawn directly from all the different cell types that make up the human body.

This type of epigenomic information has now been gathered, in the Roadmap Epigenomics Project directed by the US National Institutes of Health. This project set out to generate and publicly share epigenomic data from stem cells, from mature cells from a variety of different tissues from healthy people, and from patients with diseases such as cancer, and neurodegenerative and autoimmune disease.

The main results of this vast project are published in this issue

starting on page 313, as well as in several other Nature Publishing Group journals.

Insights into three fundamental aspects of epigenetics emerge: how the epigenome affects gene expression; how the epigenome changes during stem-cell differentiation (that is, during normal development); and how it changes during disease.

The results emphasize the central role of epigenomic information in understanding these processes. Crucially, what emerges is that it is not just one or two types of modification that matter. Biology is rarely that simple. Instead, combinations of modifications predict gene activity in ways that a single type of modification does not.

A causal link between epigenetic changes and disease has so far been hard to establish. Identifying such changes is necessary, however, if we are to understand the underlying disease mechanism and design targeted

treatments. With the new wealth of data, consistent alteration in the epigenetic landscape could identify candidate genes and pathways for further follow-up. And time-course studies of the epigenetics of cell types relevant to a specific disease could indicate whether epigenetic changes have a role in disease progression, or only in its onset.

One reason that it has been difficult to relate some diseases to disruption in DNA function is that many of the key changes occur in poorly understood regions of the genome, usually outside those parts that code for proteins. Epigenomic maps such as those published today should help scientists to navigate this poorly charted landscape. By overlaying these maps, made in relevant cell types, researchers can determine, for example, whether an epigenetic change associated with a given disease lies in a region of the genome that regulates gene activity. If it does, then this overlap provides a possible lead to be explored.

Cancer is often called the disease of the genome, but the genome does not exist, or operate, in splendid isolation. Of all diseases, cancer has been linked most unambiguously to epigenetic aberrations. Scientists have long suspected that epigenomic organization affects the genomic location of the mutations that provoke cancer. The new findings suggest that this is true, and they go further. They show that the epigenome of a cancer cell carries a fingerprint of the cell type that originated the cancer. This is crucial information, especially for cancers in complex tissues such as the liver, which cannot presently be traced to their original cell type.

In human diseases, the genome and epigenome operate together. Tackling disease using information on the genome alone has been like trying to work with one hand tied behind the back. The new trove of epigenomic data frees the other hand. It will not provide all the answers. But it could help researchers decide which questions to ask. ■

“Tackling disease using information on the genome alone has been like trying to work with one hand tied behind the back.”

The idea factory

Science will benefit most from a combination of youthful innovation and hard-won experience.

These are confusing times for senior scientists in India. Those who read the *Hindustan Times* would last month have seen an encouraging message from the science and technology ministry. A draft note sent to all government ministries, the newspaper said, proposed raising the retirement age of government-employed researchers from 60 to 65. Scientists beyond 60, it said, are still productive and contribute to the scientific wealth of the nation. Most encouragingly, it claimed, the global average age of “top scientists” is 70.

Yet, Indian Prime Minister Narendra Modi seems to have a different agenda. Late last year, Modi’s office refused to grant permission for four lab directors at the Defence Research and Development Organisation to have their contracts extended past the usual retirement age. Modi, commentators say, wants to encourage young blood and fresh talent. He wants five of the research organization’s

laboratories — including those that work on metallurgy, lasers and cryptography — to be headed by scientists aged 35 or under. At present, many young researchers go elsewhere, discouraged by the lack of opportunities in an organization dominated by the older generation.

Such bench-blocking is a problem for scientific organizations across the world. Last week, *Nature* reported on an initiative from the US National Institutes of Health (NIH), which for years has watched the average age of its grantees creep upwards (see *Nature* 518, 146–147; 2015). The NIH has proposed a system of emeritus grants that will pay senior scientists to wrap up their research and close their labs, thereby freeing up money for the next generation.

If the young scientists waiting for their turn at the top table are growing impatient, then a study suggests that they have a strong case. As we report on page 283, analysis of some 20 million biomedical papers published over the past 70 years suggests that younger researchers are more likely than older researchers to be working on innovative topics. Out with the old? Not so fast: if you are good enough then you are old enough, certainly. But the latest analysis also suggests that the most productive groups teamed a young researcher with an old(er) hand. There is an age-old problem here, but it is not necessarily old age. ■

ANNOUNCEMENT

Nature journals offer double-blind review

Starting in March, *Nature* and the monthly *Nature* research journals will offer an alternative to conventional peer review. Authors will be able to request that their names and affiliations are withheld from reviewers of their papers — a form of peer review known as double blind. At present, the process is single blind: reviewers are anonymous, but they know the authors’ identities.

Alternatives to the conventional peer-review process are often proposed. Some have suggested fully open reviews, in which the names of both authors and reviewers are known. Proponents of open peer review see its transparency as a way to encourage more civil and thoughtful reviewer comments — although others are concerned that it promotes a less critical attitude.

By contrast, advocates of double-blind peer review suggest that it eliminates personal biases, such as those based on gender, seniority, reputation and affiliation.

Both systems are already in use across scholarly publishing, but there is no consensus on which is best. *Nature* experimented with open peer review in 2006, but at the time, despite expressed interest, the uptake from both authors and reviewers was low and the open reviews were not technically substantive. Views about open peer review are probably still evolving as several journals continue to experiment with variations on this practice. Opinions about double-blind review, however, are remarkably consistent.

In one of the largest studies on peer review — a 2009 international and cross-disciplinary survey of more than 4,000 researchers — 76% of respondents indicated that double blind was an effective peer-review system (A. Mulligan, L. Hall and E. Raphael *J. Am. Soc. Inf. Sci. Technol.* 64, 132–161; 2013). (By comparison, open and single-blind peer review were considered effective by 20% and 45% of respondents, respectively.) Our own surveys confirm that double-blind peer review is a popular option. Importantly, this sentiment is widely echoed in conversations that our editors have had with young scientists worldwide. These conversations

demonstrate a widespread perception that biases based on authorship affect single-blind peer review.

The decision to offer double-blind review has been much discussed. Editors of *Nature* journals have previously resisted it for several reasons. Some were sceptical of its efficacy, some were concerned about the potential difficulty of recruiting referees, and some still saw it as their responsibility to mitigate the biases that this method tackles.

All editors take this responsibility seriously and will continue to select reviewers carefully and consider their comments. They will also continue to honour reasonable requests from authors to exclude particular reviewers, regardless of the chosen method of peer review. But by definition, unconscious biases may be difficult to identify and to control. Several studies have detected involuntary biases, notably based on gender, in other areas of the scientific enterprise, such as in the hiring of laboratory staff, citation habits and speaker line-ups at conferences.

Since June 2013, *Nature Geoscience* and *Nature Climate Change* have allowed authors to choose between double-blind and single-blind peer review at submission. The uptake of the double-blind method has been much lower than the enthusiasm expressed in surveys suggested it would be. No more than one-fifth of monthly submissions to these journals are choosing the double-blind route. No substantial effects on the quality of reviews have been detected. The positive reactions to the trial from surveyed authors are a big part in the decision to start offering double-blind review at *Nature* and the *Nature* monthly journals as well. (*Nature Communications* will join later.)

How will it work? The responsibility to render the manuscript anonymous falls to the authors. Clearly, keeping their identities from reviewers will not always be possible, especially in small and specialist fields. We also continue to promote policies that support researchers who wish to release data early and to discuss their work with their peers before publication, through conferences or by posting research on preprint servers. These routes to publication also compromise anonymity. That is why the double-blind process is optional on all titles. We expect that some authors will choose it because of concern about biases, others purely on principle.

We will keep this initiative under review, and we, of course, welcome comments from authors and reviewers. ■



Europe needs fresh focus on big-science projects

Messy governance and a lack of long-term planning threaten the success of the European Spallation Source, says Olof Hallonsten.

Big science has come to Sweden. The frozen ground near Lund, in the country's south, is being dug out to make way for Europe's latest megaproject. The European Spallation Source (ESS) is a €1.8-billion (US\$2-billion) neutron-beam machine designed to study materials structure and is scheduled to open in 2019.

The project is under way, but its future is far from secure. The funding is incomplete, the politics that support it are unpredictable and the legal framework is a mess. Yet Europe considers such risky circumstances as normal. Every collaborative big-science project in Europe has been birthed in similar messy, ad hoc circumstances.

Still, the ESS is a risk too far. If the project is not to founder, Europe — not least the two hosts, Sweden and Denmark — must learn the lessons of the past. The project must immediately be put on a more solid footing to help it through the almost inevitable cost overruns and delays, which could otherwise threaten its success and drain money from the rest of Swedish science.

Whereas the United States, Japan and others tend to run big-science projects as an arm of central government, either through federal funding or national agencies, European collaborative efforts lack this type of political grounding. The European Commission plays only a minor part — coordinating the early stages of big basic-science projects such as the ESS.

Developing collaborative projects in this way — as for the European Southern Observatory and the European Synchrotron Radiation Facility (ESRF), for example — avoids the sluggish and notorious bureaucracy of Brussels and the European Commission. But the downsides of this approach are the deals done behind closed doors, unaccountability and significant inherent uncertainty.

A brief history of the ESS demonstrates this. Germany and the United Kingdom, Europe's neutron-scattering strongholds, initially proposed to host the lab, but funds could not be agreed, and preparations ground to a halt in 2002. Even when Sweden received widespread support from other countries for its bid in May 2009, no central funding agreement was put in place. Instead, the Swedish government entered into several parallel bilateral negotiations, which still seem not to be fully resolved.

The funding solution presented by the Swedish government in July last year is so far not backed by binding agreements from all the expected contributing countries. This raises the question of whether the project will meet design specifications and scientific expectations, and at what cost. Several members have not yet progressed beyond the informal 'letters of intent' level, and so most legal and financial issues remain unresolved. Meanwhile, the governments of Sweden and Denmark have already

invested more than €100 million in the ESS project, and have recruited about 200 people to work on it. Swedish scientists and others with a stake in the project must pressure ministers to ensure that the funding pledges made by other countries are made legally watertight.

Sweden has near-zero experience of building and hosting big-science labs. Its research-policy system is decentralized and consensus-oriented, and possibly not suited to handing over significant sums to individual projects of this size. The Swedish government has pledged to pay just over one-third of the projected ESS construction cost, but what if that cost increases? Even firm supporters of neutron-scattering science such as the United Kingdom and Germany hesitated for five years after the 2009 site decision before making binding membership pledges. Will project partners be willing to pay more if necessary? If not, then where

will Sweden find the cash to meet the shortfall? Analysis of Swedish government spending plans for future years suggests that no contingency funds have been set aside. If the cost of the ESS skyrockets, then the Swedish government might have to cut back in other areas.

Existing government investment in the ESS has been financed through a complicated set of funding flows, and the numbers do not always seem to add up. This, too, signals a lack of a long-term plan, and little preparation for unforeseen events and cost increases (see O. Hallonsten *Sci. Public Policy* <http://doi.org/z8m>; 2014). Sweden must include a contingency margin in its budget. This has worked previously to minimize risk, for example with the construction of the ESRF in Grenoble, France.

Research policy is always a game of priorities, but big-science projects carry complex risks that must be properly prepared for and managed. Although the European Commission has made some moves to explore how it could establish legal frameworks for such collaborations, as well as helping to plan and set political priorities to make them happen, it is too early to predict the outcome of these efforts. It is unlikely, anyway, that new policies will be put in place in time to benefit the ESS.

The project has already suffered from the indecision of Europe in collaborative big science: while Europe has been discussing and haggling over construction and costs of the ESS, both Japan and the United States have swiftly built and started to operate their own versions. Europe is now playing catch-up. If it is not to fall further behind, then its attitude to big-science projects must change. The current preparations for the ESS are a good place to start. ■

Olof Hallonsten is a sociologist of science at Lund University in Sweden.
e-mail: olof.hallonsten@fek.lu.se

**BIG-SCIENCE
PROJECTS CARRY
COMPLEX
RISKS
THAT MUST BE
PROPERLY
PREPARED
FOR.**

➔ **NATURE.COM**
Discuss this article
online at:
go.nature.com/wvbowy

RESEARCH HIGHLIGHTS

Selections from the
scientific literature

INFORMATION TECHNOLOGY

Long-term data storage in DNA

A DNA-based system could safely store data for millennia.

Today's digital systems can store information for only around 50 years, but encoding it in DNA could greatly extend its lifetime. Robert Grass at the Swiss Federal Institute of Technology in Zurich and his colleagues have devised a system that encapsulates and protects DNA strands in silica glass. The team also included redundancy codes to correct errors that arise when writing, storing and reading the data.

Using the technique, the authors recovered 83 kilobytes of data — including the full Swiss Federal Charter from 1291 — by sequencing nearly 5,000 pieces of DNA that were kept under conditions simulating storage at around 10°C for 2,000 years.

Angew. Chem. Int. Edn <http://doi.org/f23gmf> (2015)

EVOLUTION

Fern hybrid does not mind the gap

Two ferns that last shared an ancestor more than 60 million years ago have interbred — showing that this can still happen even after a long evolutionary gap.

As populations separate and evolve over time, they lose the



ability to cross-breed. So Carl Rothfels, now at the University of California, Berkeley, and his team were surprised to find a fern (*×Cystocarpium roskamianum*; pictured) from the French Pyrenees that is a hybrid of *Gymnocarpium* and *Cystopteris*, two dissimilar genera. DNA analysis showed that its two parents diverged roughly 60 million years ago, the biggest known evolutionary gap in a plant or animal hybridization. This is comparable to a human interbreeding with a lemur. The findings suggest that

new species of fern evolve more slowly than many other plants, in part because they rely on wind and water for fertilization, making it harder for eggs and sperm of different species to remain separate.

Am. Nat. 185, 433–442 (2015)

AGRICULTURAL ECONOMICS

Trade disruptions hit the poor

Countries that cut back on food trade to protect against domestic price fluctuations can disrupt the global food system

— a sign of the increasing connectedness of the market.

Michael Puma of Columbia University in New York and his team used data on wheat and rice agriculture from 1992 to 2009 to analyse how price shocks resulting from large-scale weather anomalies, crop diseases or war affect worldwide trade in staple foods. They found that the global market has become more vulnerable to temporary trade restrictions as international connections have doubled and the volume of traded goods has increased since 1992.



PALAEONTOLOGY

Ancient mammals displayed diversity

Two fossils show that early mammals had a more varied anatomy and behaviour than was thought.

A team led by Zhe-Xi Luo at the University of Chicago, Illinois, and Qing-Jin Meng at the Beijing Museum of Natural History analysed a 160-million-year-old fossil from China.

The creature, *Docofossor brachydactylus*, had short, wide digits for burrowing underground, similar to those seen in moles. Some of the team

members studied another fossilized mammal of about 165 million years old, *Agilodocon scansorius* (pictured; US cent shown for scale). It seems to have been adapted to tree-climbing, and its teeth bear hallmarks of a diet of tree gum and sap.

The two species, which belong to an extinct group called docodonts, show that the earliest mammals lived in diverse habitats, the team says.

Science 347, 760–764; 764–768 (2015)

ZHE-XI LUO, UNIV. CHICAGO

HARRY C. ROSKAM

Export restrictions lead to higher global food prices, which can lead to more trade restrictions. Poor countries suffer most from the drop in food imports, the authors note. *Environ. Res. Lett.* 10, 024007 (2015)

PALAEOHYDROLOGY

Drying lakes linked to extinctions

Climate change in Australia may have played a part in the extinction of many large animals some 50,000 years ago.

The cause of the mass die-off is debated, with some saying that ecological collapse was sparked by human use of fire 40,000 to 60,000 years ago. Climate-related factors had been dismissed because there seemed to be little change in Australia's climate at that time. However, Tim Cohen of the University of Wollongong in Australia and his colleagues looked at sediments along the shores of two huge lakes, Eyre and Frome, and found that their water levels decreased drastically around the time when megafauna went extinct.

Lakes that shrank under a changing climate could have led to the demise of plants and herbivorous animals, the authors say.

Geology <http://doi.org/z8n> (2015)

ECOLOGY

Traps target tricky toads

Habitats that attract invasive species can be turned into 'ecological traps' that wipe out the invaders.

In Australia, invasive cane toads (*Rhinella marina*; pictured) are devastating native wildlife, and they have proved difficult to eradicate. To survive the dry season, the toads flock to ponds that store water for livestock, and then use these 'invasion hubs' as staging posts to invade more

areas during the rains.

To trap the toads, Mike Letnic at the University of New South Wales in Sydney and his colleagues used fences to exclude them from the ponds in Australia's Northern Territory.

Toads that were attracted to the water but unable to access it died in their hundreds at the fenced sites, and populations remained suppressed a year later. The authors suggest that other species that rely on invasion hubs could be controlled in a similar way.

J. Appl. Ecol. <http://doi.org/z8p> (2015)

NEUROSCIENCE

Breathe in to boost brain-fluid flow

An inwards breath drives the flow of fluid that bathes the human brain.

Cerebrospinal fluid cushions the brain, flushes out waste and in rodents seems to be controlled by pulsating blood flow. To find out how the fluid is regulated in humans, Steffi Dreha-Kulaczewski at the University Medical Center Göttingen in Germany and her colleagues used real-time magnetic resonance imaging to scan the brains of ten healthy volunteers while they did breathing exercises. The researchers found that an intake of breath had a stronger effect on fluid flow than the heartbeat did.

The approach could be used to study disorders that result in disruptions to the flow of cerebrospinal fluid.

J. Neuro. 35, 2485–2491 (2015)

PARTICLE PHYSICS

New particles found at collider

High-energy collisions between protons have unearthed two new particles.

SOCIAL SELECTION

Popular articles on social media

Science in 200 words or less

Even in this age of texts, tweets and sound bites, most scientific papers remain long and dense. But a new online journal promises to bring a little brevity to science by accepting submissions of 200 words or less. *The Journal of Brief Ideas* (<http://beta.briefideas.org>) has published only a few papers so far, but has already generated a buzz on social media. Katie Mack, an astrophysicist at Melbourne University in Australia, urged her many Twitter followers to check it out, noting that

➔ **NATURE.COM**
For more on popular papers:
go.nature.com/uzrwqb

the journal was effectively "reducing the minimum publishable unit to 200 words". But she also cautions that it could turn into a collection of preliminary ideas that are not ready for scientific consumption.

Named Ξ_b^- and Ξ_b^{*-} , the particles were discovered by the LHCb experiment team at the Large Hadron Collider at CERN, Europe's particle-physics laboratory near Geneva, Switzerland. Like protons, the particles are made up of three quarks, but they each include a heavyweight 'beauty' or 'bottom' quark, making the particles six times heavier than the proton. Although consistent with the standard model, these particular arrangements of quarks have never previously been observed.

Studying the properties of these particles could help scientists to better understand the strong nuclear forces that bind protons and neutrons together in an atom.

Phys. Rev. Lett. 114, 062004 (2015)

OCEANOGRAPHY

Arctic ice warms from below

Shrinking Arctic sea ice could cause more-vigorous mixing of ocean heat in northern waters, eventually leading to further melting.

Tom Rippeth of Bangor

University, UK, and his colleagues measured water temperatures at different depths and locations across the Arctic Ocean. They found that heat rose more quickly from warm, deep layers of water that ran into rough patches on the sea floor than from areas that have a more even floor.

Such mixing might become more common in a warming world, the authors say. As sea ice disappears, the atmosphere can transfer more of its energy into the ocean, which drives ocean mixing. The rising heat from this mixing could cause sea ice to decline even more.

Nature Geosci. <http://dx.doi.org/10.1038/ngeo2350> (2015)

CORRECTION

The Research Highlight 'Capsules collect carbon dioxide' (*Nature* 518, 140; 2015) stated that all authors are at Harvard University. In fact, Jennifer Lewis and a co-author are at Harvard; her collaborators are at Lawrence Livermore National Laboratory, California, and at the University of Illinois at Urbana-Champaign.

➔ **NATURE.COM**

For the latest research published by Nature visit:

www.nature.com/latestresearch



SEVEN DAYS

The news in brief

EVENTS

Spaceplane returns

The European Space Agency (ESA) successfully tested its prototype reusable spaceplane on 11 February, in a mission that took the craft 413 kilometres above Earth. The Intermediate eXperimental Vehicle (IXV) splashed down in the Pacific Ocean after a 100-minute trip from Europe's Spaceport in French Guiana. The IXV is trialling technology for PRIDE (Programme for Reusable In-Orbit Demonstrator for Europe), a future ESA craft that could eventually allow the agency to ferry planetary samples and humans, or to service satellites and perform low-orbit experiments. See go.nature.com/y2lgph for more.

Disease renamed

Chronic fatigue syndrome (CFS) has been given a new name by an influential US panel. In a 10 February report, the US Institute of Medicine controversially proposed the name systemic exertion intolerance disease (SEID), along with a new definition. To be diagnosed with SEID, a person must have unrefreshing sleep, fatigue that impacts life and exhaustion after any exertion. Previously, CFS was a catch-all diagnosis for patients with fatigue-related symptoms, once other diagnoses had been ruled out. The report also stated that SEID is a physiological, not psychological, disease. See go.nature.com/eutffs for more.

Deep-space probe

The Deep Space Climate Observatory (DSCOVR) soared into space on 11 February, ending more than a decade of battles over the Earth- and space-weather-observing probe.



KARL MONDON/TNS/NEWSCOM

US flight body drafts looser drone rules

The US Federal Aviation Administration (FAA) has proposed regulations that would more freely allow 'commercial' drone flights, which include those made for research purposes (pictured). Such flights are currently not permitted without FAA approval, which is given on a

case-by-case basis. Under the draft rules, released on 15 February, drones weighing less than 25 kilograms would be allowed to fly in daylight at up to 161 kilometres per hour and a maximum altitude of 152 metres, provided that they remain in view of the person operating them.

A Falcon 9 rocket from the private company SpaceX lifted off from Cape Canaveral, Florida, carrying the National Oceanic and Atmospheric Administration spacecraft. DSCOVR, originally planned as an Earth-imaging satellite by former US vice-president Al Gore, will monitor space weather from about 1.5 million kilometres above Earth. It was the first deep-space launch by SpaceX, which, because of rough seas, cancelled an opportunity to try to land the rocket's used first stage on a barge.

Japan aftershock

An earthquake shook the northeast of Japan on 17 February. The quake's magnitude was recorded by the Japan Meteorological

Agency (JMA) as 6.9 and by the US Geological Survey as 6.7. According to media reports, a seismologist from the JMA told reporters that the event was an aftershock of the 2011 Tohoku quake that caused huge damage on Japan's coast.

POLICY

Wildlife trade

The United States has outlined how it will increase efforts to stamp out illegal wildlife trading. The departments of justice, state and the interior say that they will strengthen controls over US trade in ivory, among other measures, and reduce demand for certain wildlife products. The plan, released on 11 February, builds on President Barack Obama's 2014 strategy for combating

the wildlife trade. Meanwhile, in South Africa, environment minister Edna Molewa announced on 10 February that a committee has been established to investigate the feasibility of a regulated trade in rhino horn. Molewa said that any proposal would be "based on sound research".

Risky climate plans

More research is needed into some areas of the nascent field of geoengineering, say a pair of reports by the US National Research Council. The findings, published on 10 February, could set the stage for a formal US research programme to test ways to manipulate Earth's climate. The reports argue that the risks of schemes that aim to alter Earth's reflectivity to

COURTESY ANU MEDIA

cool it down often outweigh the benefits. Small-scale research should begin on ways to capture and store carbon permanently, the authors say. See go.nature.com/ddxd2j for more.

Climate draft

A draft text for negotiation in the run-up to the next big United Nations climate meeting was hammered out last week. After six days of talks in Geneva, Switzerland, delegates came up with the 86-page document as the basis for the climate agreement to be finalized by nations in Paris later this year. That accord will come into force in 2020, and will aim to keep average global temperatures from rising by more than 2°C.

PEOPLE

Climate chief dies

Michael Raupach, a leading Australian climate and ocean scientist, died on 10 February, aged 64. After a long career in marine and atmospheric research at the Commonwealth Scientific and Industrial Research Organisation (CSIRO) in Canberra, he was last year appointed director of the Australian National University's Climate Change Institute in Canberra. Raupach (pictured) was best known for his work in atmospheric physics and on water and



carbon cycles. In 2001, he co-founded the Global Carbon Project, an initiative to establish a knowledge base to inform policies to curb greenhouse-gas emissions.

RESEARCH

Grey-wolf woe

A grey wolf spotted near the Grand Canyon in October last year — the first in the area since the 1940s — was shot and killed in Utah in December, according to a genetic analysis by the US Fish and Wildlife Service. The results were announced on 11 February. The three-year-old female is thought to have travelled at least 1,200 kilometres from Wyoming, where it had been tagged with a radio collar in January 2014. The animal's extensive wandering attracted so much interest that a national contest was held to give her a name, Echo. Wolves in Utah are protected under the Endangered Species Act; the

hunter allegedly mistook the animal for a coyote.

Ebola report

Clinical trials of Ebola vaccines should continue in West Africa — even if there are too few new cases to determine definitively whether they work, urges a 17 February report. Released by UK biomedical charity the Wellcome Trust and the Center for Infectious Disease Research and Policy at the University of Minnesota, Minneapolis, the report says that testing and manufacturing of Ebola vaccines and medicines should be ramped up in case there is a resurgence of the current outbreak in West Africa, and to combat future Ebola epidemics.

FACILITIES

Ultrafast Arab laser

The Arab world's first attosecond-laser facility opened at the King Saud University in Riyadh on 16 February. Such facilities use laser pulses lasting a few billionths of a billionth of a second to capture images of the motion of electrons for study. In collaboration with the Max Planck Institute for Quantum Optics in Garching, Germany, and the University of Munich, the Saudi Arabian facility will carry out research in atomic

COMING UP

19 FEBRUARY

The US National Academy of Engineering will launch a contest called The Next MacGyver, crowdsourcing ideas for a television show featuring a female engineer as the lead character. The competition, created in response to a White House request, involves Hollywood producers including the creators of the popular series *MacGyver* and *CSI*. go.nature.com/wz8ban

23–27 FEBRUARY

The Kavli Institute for Theoretical Physics in Santa Barbara, California, will hold a meeting on the physics of exoplanets, with particular emphasis on 'super-Earths' — planets with radii 2.5 times that of Earth. go.nature.com/jcsuf6

physics and molecular biology. See page 281 for more.

BUSINESS

Transgenic apples

Apples genetically engineered to be 'non-browning' when cut have been deregulated by the US Department of Agriculture, meaning that they can be grown and sold to consumers. Arctic Apples were developed by Okanagan Specialty Fruits in Summerland, Canada, to have reduced levels of an enzyme responsible for turning fruit flesh brown when it is exposed to air. This is one of the first genetically modified plant products with a trait aimed at consumers rather than farmers. The first of the apples could be on the market by 2016.

➔ NATURE.COM

For daily news updates see: www.nature.com/news

SOURCE: HESA

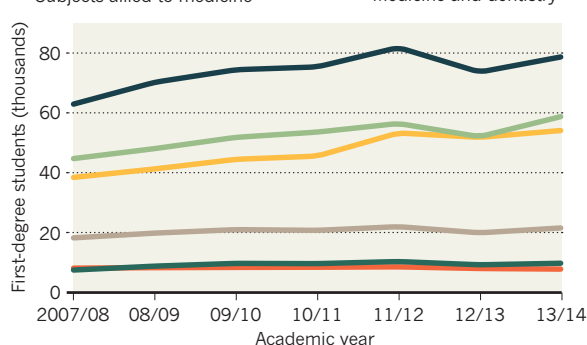
TREND WATCH

Some science subjects have seen large hikes in undergraduate student numbers in Britain over the past seven years, according to a report by the UK Higher Education Statistics Agency. The biggest winners were subjects allied to medicine, such as nursing and pharmacology, which grew by 39%, followed by biological sciences, which rose by 30%. The next highest gains in popularity were jointly in maths and in business and administrative studies, which attracted 24% more students.

BIOMEDICAL-SCIENCES BOOM

UK students have turned to subjects allied to medicine, biology and maths in the past seven years.

■ Business and administrative studies ■ Physical sciences
■ Biological sciences ■ Mathematics
■ Subjects allied to medicine ■ Medicine and dentistry



NEWS IN FOCUS

DATA-MINING Young scientists are the most open to new ideas **p.283**

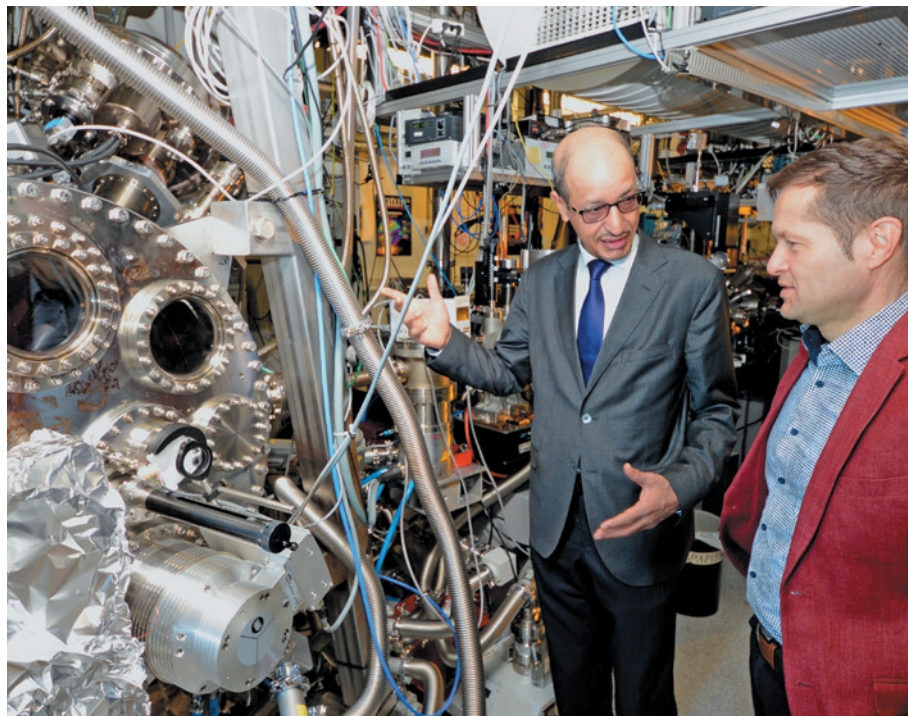
LINGUISTICS Eastern steppe tipped as seat of Indo-European languages **p.284**

BIODEFENCE Miniature human organs grown on plastic chips **p.286**

TAXONOMY Collections grow in value even as they fall into decline **p.292**



THORSTEN NAESER



Abdallah Azzeer (left) and Ferenc Krausz head up a collaboration at King Saud University in Riyadh.

ARAB SCIENCE

Saudi Arabia opens top-notch laser lab

Facility will study biomedical uses at attosecond resolution.

BY ALISON ABBOTT

A cutting-edge laser facility — the first of its kind in the Arab world — opened this week at Saudi Arabia's oldest and largest university. The launch pushes forward the country's ambitions to become a leader in science and builds on a collaboration with Western scientists that has required some cultural adjustments.

The Attosecond Science Laboratory at King Saud University (KSU) in Riyadh hosts an 'atto-second laser', which generates ultrashort pulses of light, lasting just a few billionths of a billionth

of a second, that can image otherwise invisible electrons as they move similarly fast within atoms. Attosecond lasers were invented in 2001, and facilities now exist at dozens of sites around the world. The Saudi Arabian facility is the result of a collaboration that began in 2008 with the Max Planck Institute of Quantum Optics (MPQ) in Garching, Germany, which hosts its own attosecond laser, and the Ludwig Maximilian University of Munich.

"It is very exciting that the frontier of attosecond science is now having its outpost in the Gulf state," says Olga Smirnova, an atomic physicist at the Max Born Institute in Berlin.

Saudi Arabia is known for its oil wealth, and in 2002 its government decided that science was the key to incubating a more diverse economy. Its strategy comprises heavy financial investment and forging partnerships with leading research institutions abroad — and it seems to be working. In the past five years, the number of scientific papers produced by researchers in Saudi Arabia has skyrocketed. The quality of the research has now overtaken that of Turkey and Iran, according to impact metrics known as SNIP (Source Normalized Impact per Paper) from the University of Leiden in the Netherlands. Prince Turki Bin Saud Bin Mohammad Al Saud, who heads Saudi Arabia's National Science, Technology and Innovation Plan, told *Nature's* News team that science funding has been doubled from this year and that the country is on track to reach Western levels by the mid-2020s.

Attosecond lasers quickly became fundamental tools in atomic physics after the first attosecond laser pulses were reported in 2001 by a team led by the MPQ's Ferenc Krausz, who heads the Attosecond Science Lab (M. Hentschel *et al. Nature* **414**, 509–513; 2001).

The lasers have since moved into the realm of molecular sciences, including condensed-matter systems and molecular biology, where they are being used to investigate how the movement of electrons can initiate changes in the structure of molecules. "They provide an exquisitely sharp temporal scalpel for dissecting the inner workings of matter," says laser physicist John Tisch of Imperial College London.

One of the first planned experiments for the KSU laser will study the behaviour of electrons in atoms of melanin, best known as the pigment that protects skin from the sun's ultraviolet rays. No one knows why ultraviolet photons do not normally break the chemical bonds in the molecule when they hit it, but it is assumed that melanin's electrons redistribute — and diffuse — the energy among themselves. The experiment at KSU will test this hypothesis by developing extremely short, high-intensity ultraviolet pulses to excite the electrons, and will then capture their movements with the attosecond laser.

The collaboration with Saudi Arabia gives Krausz the chance to enter totally new territory. He will work with oncologist Jean-Marc Nabholz — who moved to KSU last year to head its National Comprehensive Cancer Center — to adapt the laser to generate ▶

► pulses of infrared light for analysing proteins and nucleic acids in blood samples from people with cancer. The aim will be to find molecular ‘fingerprints’ that might diagnose cancers, or predict response to therapy or the future onset of a cancer.

The value of such a source of infrared light, Krausz says, is that a table-top-size laser system could be developed and used at patients’ bedsides. Currently, the only sources of such radiation are synchrotrons, which require large, expensive infrastructures. Because Krausz has little experience in this area, it would have been hard for him to obtain funding in Germany for such medical applications, he says.

The place of women in Saudi society and education, and the country’s human-rights record, have presented challenges for members of the collaboration. At KSU, which was founded in 1957, male and female students have separate campuses. No rule forbids women from entering the new lab, says Abdallah Azzeer, who leads the KSU side of the laser collaboration, but mixing of the sexes contravenes cultural norms. “We will make special arrangements to ensure their access,” he says. One possibility, he adds, might be to train female PhD students in handling the equipment so that they can supervise female undergraduates whose parents do not want them to attend mixed classes.

Krausz has had to get used to working in a segregated environment during his time at KSU. All the lectures are given in the men’s campus and beamed over to the women’s campus, and Krausz remembers being extremely startled the first time he received a disembodied question from a female student over loudspeakers.

He thought long and hard about working with Saudi Arabia, he says. As a Hungarian who left for the West in 1987 at the age of 25, he is hypersensitive to human-rights issues. But not long before he decided to collaborate with KSU in 2008, he had cancelled a trip to China in protest against a clamp-down on press freedom there, and then regretted the decision. It achieved nothing save the embarrassment of the scientists, he says, and he concluded that, in such cases, “the best thing is to talk to each other and learn each other’s problems”.

He found himself genuinely moved by the enthusiasm for science he encountered on his first visit to KSU later that year. “It felt like a small revolution was happening,” he says. “I thought about how I would have felt in the same situation in Hungary — I might have stayed.” ■

“They provide an exquisitely sharp temporal scalpel for dissecting the inner workings of matter.”



US President Barack Obama has ordered better federal cooperation with private firms to fight hackers.

CYBERSECURITY

Cybercrime fight targets user error

Researchers consider ways to diminish human factors in the equation for keeping data safe.

BY ERIKA CHECK HAYDEN

It would be easy to blame the poor soul at Sony Pictures Entertainment who opened the door to one of the most disastrous hacks in history just by clicking an e-mail link. As US President Barack Obama pointed out during a visit to Stanford University in California on 13 February, user negligence is often the key to a successful cyberattack.

“It’s just too easy for hackers to figure out usernames and passwords, like ‘password’. Or 12345 ... 7,” Obama said. But people do this kind of thing all the time, says Angela Sasse, head of information-security research at University College London. Researchers have found that after employees were asked to create long passwords according to strict rules, some of them wrote the password down in an easily accessible place, such as on a desk in plain sight. Other employees might choose to work outside a secured network because it runs too slowly (see also go.nature.com/buxsds).

Such measures confound security experts but are a logical response to the increasing

security workload imposed on employees, Sasse says. “We want security that is effective but also allows us to get on with the job,” she adds. “A lot of smarter companies are realizing that some of these security measures are a bad productivity drain.” Cormac Herley, a security researcher at Microsoft Research in Redmond, Washington, has estimated that the world’s Internet users collectively spend the equivalent of 1,389 years every day entering passwords (C. Herley *IEEE Secur. Priv.* **12**, 14–19; 2014).

Generally, the financial services industry is further ahead than others in dealing with the problem, because its business relies on ensuring that customers can easily access their funds while thieves are kept out. But a spectacular failure in its efforts was revealed on 16 February, when the Russian computer-security firm Kaspersky Lab described how hackers had managed to steal an estimated US\$1 billion from financial institutions around the world by infiltrating a bank in Ukraine. As in the Sony case and many others, the fatal security flaw was an errant click on an e-mailed link.

In banking, authentication is the key step

NICHOLAS KAMM/AFP/GETTY

— verifying that someone trying to access funds in a customer's name is the actual customer. This is increasingly done through layers of multiple passwords that must meet rules on length and complexity, making them hard to enter correctly on mobile devices, for example.

Some banks are experimenting with ways to jettison passwords altogether. In 2013, major German banks deployed a system called photoTAN that uses an application downloaded to a phone or desktop computer to ensure that only customers can see e-mailed account information and that hackers cannot send counterfeit e-mails. The system mathematically encodes transaction information into an image that looks to a hacker or any other observer like a meaningless jumble of coloured squares. But when a customer with the application snaps a photo of the image, it is decoded to reveal the transaction information.

A project by Google aims to revamp a system known as CAPTCHA, which distinguishes humans from programs called bots that can be used in various malicious ways, such as harvesting e-mail addresses. The existing CAPTCHA format asks a computer user to retype a line of distorted text to make the distinction, but

as artificial intelligence has advanced, the text distortion has increased such that it often defies humans as well as machines. Google's project aims to make this verification process less painful, and even invisible.

In December, Google deployed a system that, according to the company's online-security blog, "considers a user's entire engagement with the CAPTCHA — before, during, and after — to determine whether that user is a human". Google has not specified what that means, but it is believed to involve tracking a person's browser history and spotting distinctively human cues in how the cursor moves to the text box, for instance. In some cases, the program can verify that a user is human without the person even completing the task.

Another effort being spearheaded by Google, along with the file-hosting service Dropbox and the Open Technology Fund in Washington DC — an organization funded by the US government to foster free speech online — aims to improve user experience to make e-mail encryption easier. There are two existing programs, Pretty Good Privacy (PGP) and GNU Privacy Guard (GPG), which are 'open source' and so can be used by anyone to make e-mail

completely indecipherable to those who might intercept it. The systems are safe and effective: whistle-blower Edward Snowden specifically chose to leak US National Security Agency documents to documentary film-maker Laura Poitras because she uses encryption software. He knew that he could communicate with her without fear of anyone eavesdropping.

But these systems are difficult to use, so many people do not. That makes it much easier for cybercriminals to entice people to click on links, especially in the increasingly distracting online world. Google and its partners have helped to establish a non-profit online privacy consultancy called Simply Secure, which is now helping developers of such open-source programs to improve the experience for users. If the effort succeeds, the practice of using counterfeit e-mails to lure people into clicking malicious links could become much less prevalent.

But just as in conventional conflicts, the war against hackers is an arms race. "We design new defences, and then hackers and criminals design new ways to penetrate them," Obama said at Stanford. "So we've got to be just as fast and flexible and nimble in constantly evolving our defences." ■

CAREER DEVELOPMENT

Young scientists go for fresh ideas

Analysis of millions of papers finds that junior biomedical researchers tend to work on more innovative topics than their senior colleagues do.

BY EWEN CALLAWAY

Bad news, scientists: there is a good chance that your most cutting-edge work is behind you.

Young researchers are much more likely than older scientists to study exciting innovative topics, according to a text analysis of more than 20 million biomedical papers published over the past 70 years. More-senior researchers are more likely to publish in hot areas when they are supervising a younger scientist.

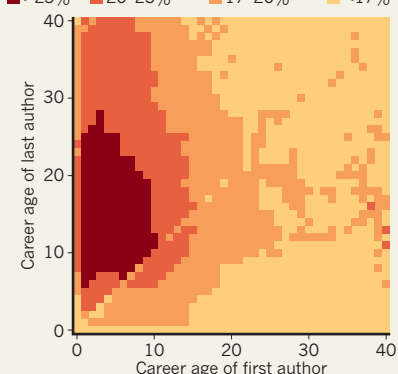
Researchers are at their most creative when they are young, or so says conventional wisdom: Charles Darwin and Max Planck both argued that young scientists were more open than older colleagues to new ideas. But the topic is not just fodder for chats over post-seminar beers. Funders such as the US National Institutes of Health have implemented policies specifically to support early-career scientists, based in part on the view that young researchers are more innovative than seasoned scientists. And in mathematics, the Fields Medal has been reserved for researchers under 40.

"It's always just a claim — the young are

HOT SPOT

Pairings of young first authors and mid-career last authors are the most likely to work on the hottest biomedical topics.

Share of publications trying out new ideas
 ■ >23% ■ 20–23% ■ 17–20% ■ <17%



more innovative — but there's no proof," says Mikko Packalen, an economist at the University of Waterloo in Canada, who led the study with economist Jay Bhattacharya of Stanford University in California. Their working paper

was published this month by the US National Bureau of Economic Research (M. Packalen and J. Bhattacharya Preprint at <http://doi.org/z87>; 2015).

To determine which scientists used the most innovative ideas, Packalen and Bhattacharya turned to the leading index of biomedical research, MEDLINE (accessed through the website PubMed), which stores more than 21 million articles published since 1946.

The duo developed a computer program that identifies every one-, two- or three-word string in the title and abstract of each paper. It then logs when each string first appeared in the literature and counts how many times it has appeared subsequently, to determine its popularity. (The all-time winning concept was 'polymerase chain reaction', the DNA-copying technique, occurring in more than 176,000 titles or abstracts.)

Packalen and Bhattacharya then ranked the most innovative articles for each year, from 1946 to 2011, on the basis of whether they were an 'early adopter' of the hottest keywords.

The method could not measure researchers' creativity, only their willingness to ▶

► embrace new ideas, which might have been proposed by others. But it showed that except for the newest scientists, young researchers far outpaced older scientists in citing new ideas in their papers, Packalen and Bhattacharya found. Because the two had no way of measuring the actual age of a researcher, they calculated 'career ages' — the number of years after a scientist's first publication.

"I really like the way they're approaching things in terms of text analysis," says Bruce Weinberg, an economist at Ohio State University in Columbus, who works with Packalen and Bhattacharya on other projects.

All is not lost for senior scientists, however. Packalen and Bhattacharya also analysed the career stages of papers' first authors (who tend to do the bulk of the research) and last authors (who tend to be supervisors), and found that the most innovative combination was an early-career first author and a mid-career last author (see 'Hot spot').

"One reading of the results is that young guys are innovative but they also need some mentorship."

was true: that young guys are innovative but they also need some mentorship," says Packalen.

And Weinberg previously found that the age at which scientists make Nobel-prizewinning breakthroughs is increasing (B. F. Jones and B. A. Weinberg *Proc. Natl Acad. Sci. USA* **108**, 18910–18914; 2011). "I think we're learning something about what these different measures are picking up," Weinberg says.

Nonetheless, Paul Ginsparg, a physicist at Cornell University in Ithaca, New York, and the founder of the online repository arXiv.org, says that Packalen and Bhattacharya's findings make sense. "In some areas of biomedical research it might take a couple of years to learn a new set of ideas and retool a lab," he says. "Hence it wouldn't be surprising if established researchers have trouble finding the time to do so."

Ginsparg also wonders whether analysing the full text of papers might tell a different story. It could be that established researchers incorporate fresh ideas into an existing methodology and framework, and therefore mention them deeper in a paper.

Packalen, who published his first paper in 2010, knows that the findings could be tough for some older scientists to swallow. "I look at these findings and say, 'No way is this going to happen to me,'" he says. "I'm going to stay innovative. I'm going to learn new ideas." ■

LINGUISTICS

Language origin debate rekindled

Eurasian steppe gains ground as Indo-European birthplace.

BY EWEN CALLAWAY

From Icelanders to Sri Lankans, some 3 billion people speak the more than 400 languages and dialects that belong to the Indo-European family. Two fresh studies — one of ancient human DNA, the other a newly constructed genealogical 'tree' of languages — point to the steppes of Ukraine and Russia as the origin of this major language family, rekindling a long-standing debate.

Scholars have long recognized an Indo-European language group that includes Germanic, Slavic and Romance languages as well as classical Sanskrit and other languages of the south Asian subcontinent. Yet the origins of this family of tongues are mired in controversy.

Some researchers hold that an early Indo-European language was spread by Middle Eastern farmers around 8,000–9,500 years ago (see 'Steppe in time'). This 'Anatolian hypothesis' is supported by well-documented migrations into Europe, where agriculturalists replaced or interbred with the existing hunter-gatherers. In 2012, a team led by evolutionary biologist Quentin Atkinson of the University of Auckland in New Zealand produced a family tree of Indo-European tongues that also pointed to an Anatolian origin more than 8,000 years ago.

A competing theory posits that the

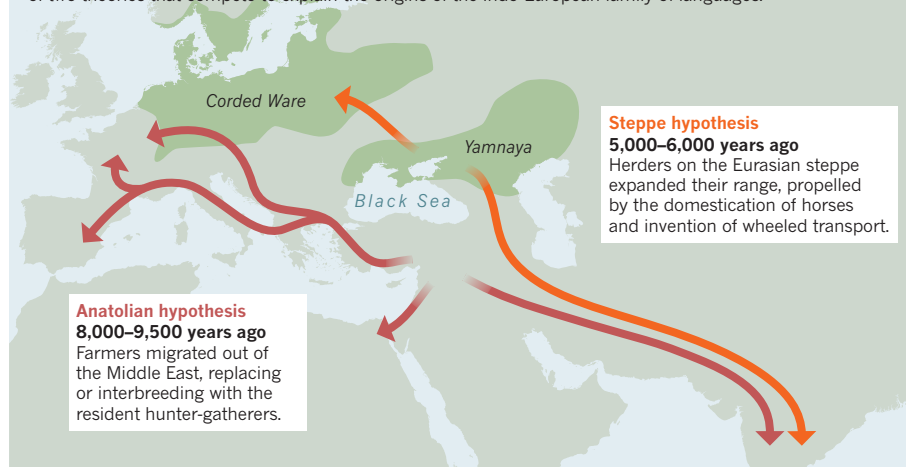
languages emerged on the Eurasian steppe some 5,000–6,000 years ago, when the domestication of horses and invention of wheeled transport would have allowed herders there to rapidly expand their range. Proponents of the 'steppe hypothesis' note that linguistic reconstructions of a proto-Indo-European tongue include words associated with wheeled vehicles, which were not invented until long after Middle Eastern farmers had reached Europe. "Most linguists have signed up to the steppe hypothesis," says Paul Heggerty, a linguist at the Max Planck Institute for Evolutionary Anthropology in Leipzig, Germany.

One knock against the theory was a lack of compelling evidence for a large-scale migration from the Eurasian steppe at this time.

A study of ancient human DNA posted to the bioRxiv.org preprint server on 10 February now plugs that gap (W. Haak *et al.* <http://doi.org/z9d>; 2015). A team led by David Reich, an evolutionary and population geneticist at Harvard Medical School in Boston, Massachusetts, analysed DNA from the bodies of 94 individuals who lived across Europe between 8,000 and 3,000 years ago. The data confirmed the arrival of Middle Eastern farmers in Europe between 8,000 and 7,000 years ago. But they also revealed evidence for a second migration that began several thousand years later. DNA

STEPPE IN TIME

An ancient-DNA study links the Corded Ware culture of northern Europe with the Yamnaya culture of the Eurasian steppe. It points to a mass migration northwest that would support the Steppe hypothesis, one of two theories that compete to explain the origins of the Indo-European family of languages.



W. HAAK ET AL. [HTTP://DOI.ORG/Z9D](http://doi.org/z9d) (2015)

recovered from steppe herders called the Yamnaya, who lived in what are now Russia and Ukraine around 5,000 years ago, closely matched that of 4,500-year-old individuals from present-day Germany, who were part of a group known as the Corded Ware culture that encompassed most of northern Europe. The similarities suggest “a massive migration into the heartland of Europe from its eastern periphery”, the team writes.

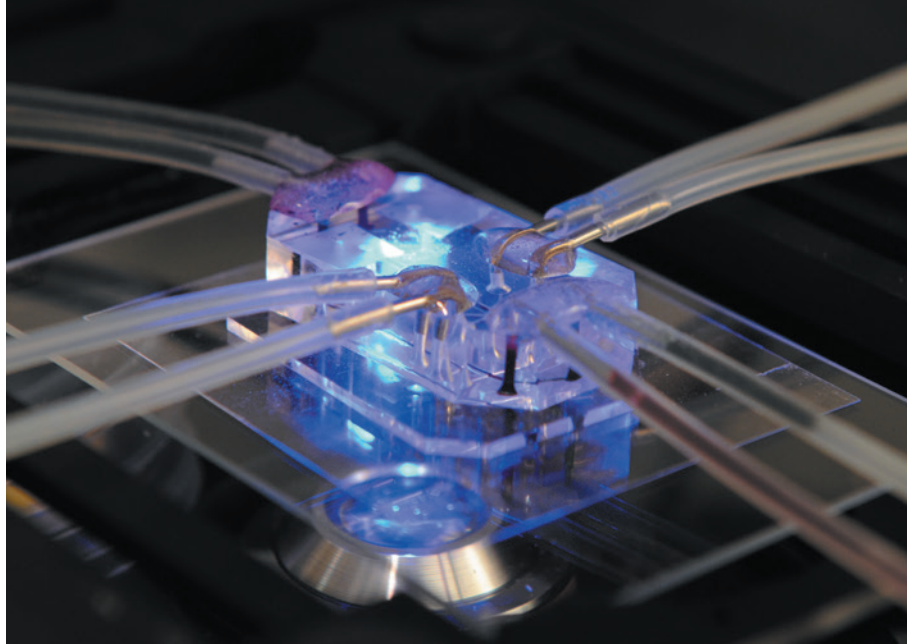
Yamnaya ancestry survives in the genomes of modern Europeans, with northerners such as Norwegians, Scots and Lithuanians maintaining the strongest link. The geographical extent of the Yamnaya migration is not clear, but the researchers note that the eastern migrants could have completely replaced existing populations, at least in what is now Germany. It is impossible to know the language these migrants spoke, but it is likely to have originated in the steppe homelands of the Yamnaya.

“This seems like very striking support for at least part of the traditional steppe model of Indo-European diversification,” says Andrew Garrett, a historical linguist at the University of California, Berkeley, whose own work adds further support. When he and his team re-analysed the data from Atkinson’s 2012 family tree, this time taking into account the approximate ages of ancient Indo-European languages, they dated the origin to around 6,000 years ago, in line with the steppe hypothesis (W. Chang, *et al. Language*; in the press).

Atkinson says, however, that the analysis assumes that ancient languages such as Latin and Old Irish are direct ancestors of modern languages, instead of side-branches of a common ancestor. This makes it appear that these languages evolved faster than they did, he says, and would argue incorrectly for a more-recent common tongue.

Heggerty points out that Reich’s ancient DNA study is not the final word on the steppe hypothesis either. He suspects that the Yamnaya spoke a language that later developed into Slavic, Germanic and other northern European tongues, but he doubts that the group imported the predecessor of all Indo-European languages: “For me, these data look like the steppe population was speaking a branch of Indo-European.”

Reich and his team acknowledge that attributing the origin of all Indo-European languages to the Yamnaya migration would require the discovery of their genetic signatures in samples from further east, such as from India and Iran. But Carles Lalueza-Fox, a palaeogeneticist at the Institute of Evolutionary Biology in Barcelona, Spain, notes that the climates of the Middle East and southern Asia do not augur well for preservation of ancient DNA: “It could be difficult to find good samples from the right time frame.” ■



‘Organs on chips’, such as this simulated lung, could be used to test bodily responses to toxic chemicals.

BIOENGINEERING

Scientists seek ‘Homo chippiens’

Biodefence projects aim to mimic the human body using networks of simulated organs.

BY SARA REARDON

Each year, the US government spends hundreds of millions of dollars stockpiling countermeasures for potential biological, chemical and radiological warfare agents. For ethical reasons, many of these treatments have never been tested in humans. Now, the US military and civilian science agencies are supporting the development of the next best thing for tests: miniature human organs on plastic chips.

“It’s unethical to expose humans to the kind of radiation that you’d see in a disaster like Fukushima, but you need to be prepared,” says Donald Ingber, a bioengineer at Harvard University’s Wyss Institute in Boston, Massachusetts. With support from the US Food and Drug Administration, he is adapting his ‘bone marrow on a chip’ to study the effects of harmful radiation and experimental remedies.

Other researchers working along similar lines discussed their work on model organs for biodefence applications at a meeting of the American Society for Microbiology (ASM) last week in Washington DC. The hope is that these complex three-dimensional systems will mimic human physiology better than do cells grown in a dish, or even animals.

A common way to form a model organ is to seed cells into channels in a small plastic chip and then feed them with nutrient-rich

fluid that flows through the system to mimic blood. The devices can be used individually or connected to other types of organs-on-chips to approximate a biological system, or — eventually — perhaps an entire human body.

The US Environmental Protection Agency plans to announce next month an US\$18-million programme to link ‘livers on chips’ with chips that simulate fetal membranes, mammary glands and developing limbs. The ultimate aim is to study how environmental contaminants such as dioxin and bisphenol A alter metabolism in those organs once they have been processed by the liver.

The flexibility afforded by model-organ systems is especially attractive to researchers who are investigating dangerous pathogens, given the expense of animal studies and the security restrictions required. At the ASM meeting, microbiologist Joshua Powell of the Pacific Northwest National Laboratory in Richland, Washington, presented experiments testing the ability of anthrax spores to infect a three-dimensional ‘lung’ grown from rabbit lung cells. The cells sit at an interface between liquid and air, much as in real lungs.

Powell says that the US Department of Homeland Security is interested in using the system to answer questions such as how many anthrax spores are necessary to cause disease in the body.

For some viruses in particular, Ingber ►

► says, researchers “have no idea about the mechanism, and they need the mechanism to get new drug targets”. Infecting model organs could allow researchers to watch how gene expression and metabolism change in real time.

This sort of information could also be used to identify an unknown agent during a chemical, biological or radiological attack, by providing baseline data on known agents for comparison. John Wikswo, a physiologist at Vanderbilt University in Nashville, Tennessee, and his colleagues have shown that they can rapidly distinguish poisons such as ricin and botulinum toxin by analysing the metabolic activity of cells (S. E. Eklund *et al.* *Sensors* **9**, 2117–2133; 2009), and will now apply the model-organs approach.

Researchers have already developed dozens of individual model organs; the next challenge is to hook them together with the eventual goal of forming an entire human body on a chip, says Kristin Fabre, a programme manager at the National Center for Advancing Translational Sciences (NCATS) in Bethesda, Maryland. This would provide a more accurate picture of the effects of a drug, toxin or other agent on human physiology.

Wikswo humorously dubs such a system *Homo chippiens* — but warns that simulating a human body will not be easy. Among other challenges, the blood substitute that flows between model organs must reach them in the right order and in the right quantity, and carry the right nutrients for each organ.

But plenty of people are trying. An NCATS-funded project aims to hook together at least 4 chips; 11 research teams are participating. The US Department of Defense’s Defense Advanced Research Projects Agency is supporting the development of techniques to link ten organs, and its Defense Threat Reduction Agency aims to build two four-organ systems.

Fabre predicts that some of the systems could be available to academics and industry within five years. She is hopeful that they will prove especially useful in cases in which animals are poor models for human physiology. As researchers inch closer to that goal, she says, “it’s like sci-fi comes to life every day.” ■



An ARPA-E-funded technology to control power flow is being used by US utility companies.

TECHNOLOGY

Radical energy ideas secure private funds

US federal start-up funds inspire investment in ARPA-E technologies.

BY JEFF TOLLEFSON

As it enters its seventh year, an ambitious US Department of Energy effort to trigger innovation in clean-energy technology is celebrating some success. At the start of the annual summit of the Advanced Research Projects Agency—Energy (ARPA-E) on 9 February, the programme’s management announced that ARPA-E-supported technologies have attracted US\$850 million in private investment.

At the same time, however, the market for new ideas in energy is not booming, and the mood at the summit was more sedate than in earlier years, when sessions swarmed with investors looking for the next big thing. Most observers of energy innovation counsel patience: even useful ideas often take decades

to reach broad application. The US defence department’s Defense Advanced Research Projects Agency (DARPA) was the model for ARPA-E, and its biggest success story — the Internet — took decades to be recognized.

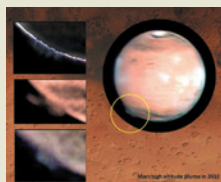
“If we look in 20 or 30 years and can’t see an impact, then we can say we have failed,” says Ilan Gur, a former ARPA-E programme director who now heads a programme to advance technological innovation at Lawrence Berkeley National Laboratory in Berkeley, California. “The real contribution that ARPA-E has had is in laying out the challenges and bringing together the communities that can solve them.”

General enthusiasm for nudging new technologies to market has shrunk since ARPA-E was founded. Venture-capital investments in the United States have dropped off sharply in the past two years; funding for early-stage

SMART WIRES


**MORE
ONLINE**

TOP NEWS



Mysterious
Martian
plume
flummoes
scientists
[go.nature.
com/ars6fw](http://go.nature.com/ars6fw)

MORE NEWS

- Jupiter tests exoplanet technique go.nature.com/gmtj3y
- United States to have worst drought in 1,000 years go.nature.com/esxz2a
- Forensic scientist builds bridges to lawyers go.nature.com/b9cxp8

Q&A



Braving
ISIS to do
science
in the
Sahara
[go.nature.
com/gbls2a](http://go.nature.com/gbls2a)

MICHAEL BAUSER

clean-energy research has almost disappeared. The administration of US President Barack Obama is seeking to reverse that trend with an initiative, announced on 10 February, to increase private investment. As part of this, the University of California Board of Regents has committed to invest \$1 billion of its endowment and pension in climate-friendly technologies. The White House also plans to hold a clean-energy investment summit in coming months.

In its six years, ARPA-E has invested around \$1.1 billion in 400 projects, ranging from large-scale electricity storage to modern grid technologies, improved biofuels and efficient methods of capturing carbon dioxide from power plants. It is often credited with injecting enthusiasm into the field. But in the slow-moving industry, its actual impact has been difficult to determine.

At the 2015 Energy Innovation Summit near Washington DC, ARPA-E deputy director Cheryl Martin described how companies that are peddling agency-supported technologies have already invested more than six times the original ARPA-E investment.

"I think we are starting to make a difference," says Martin. "We are tackling global problems and having tangible results."

ARPA-E grants are relatively large — up to \$10 million over three years — but come with strings. Projects must have ongoing, direct

engagement with programme managers, and meet strict performance benchmarks. So far, the agency has cancelled 21 projects, and staff members have worked to redesign others in light of surprising results.

The agency funds academics and small start-up companies, but it also supports research by industrial giants such as United Technologies Corporation (UTC) of Hartford, Connecticut, which has led several projects and has been involved in about a dozen others. Company officials say that these projects — which include batteries, advanced manufacturing, refrigeration and natural-gas storage — are in a grey area where the market does not justify private investment in early-stage technology.

"ARPA-E really did create an ecosystem of innovation that did not exist before," says Craig Walker, who oversees energy technology at UTC's research arm. He estimates that perhaps two-thirds of the work funded by ARPA-E over the past six years falls into categories that did not have a home in conventional funding agencies.

That is how ARPA-E was conceived. A 2007 report by the US National Academies proposed an agency to fill a gap between basic and industrial energy research. It called for ARPA-E's budget to build up to around \$1 billion annually over the course of five or six years, but despite fairly broad bipartisan support, the agency has been stuck at around

\$280 million since 2012.

For Douglas Kirkpatrick, chief executive of technology firm BlackPak of San Francisco, California, the funding situation is not necessarily a bad sign. He says that many investors entered the game a few years ago thinking

"ARPA-E really did create an ecosystem of innovation that did not exist before."

that they could get rich on a hot new energy technology, as with the Internet boom, but energy markets are slow and methodical.

BlackPak is developing a low-pressure natural-gas tank that could loosen petrol's grip on the transportation sector. With that ambitious goal, the project has attracted \$4.6 million in ARPA-E funding, but its first step is decidedly modest: corner the market in golf-course lawnmowers. Cars come later. "You start by chewing on the tail of the elephant," says Kirkpatrick.

He learned that lesson during eight years at DARPA, first as a programme manager and then as chief scientist. DARPA targets its products at the US military rather than golf courses, Kirkpatrick explains, but the same principle applies: you need patience to see whether a technology will spread.

"You've just got to populate the space and then wait." ■

SEX REDEFINED

THE IDEA OF TWO SEXES IS SIMPLISTIC.
BIOLOGISTS NOW THINK THERE IS A
WIDER SPECTRUM THAN THAT.

BY CLAIRE AINSWORTH

As a clinical geneticist, Paul James is accustomed to discussing some of the most delicate issues with his patients. But in early 2010, he found himself having a particularly awkward conversation about sex.

A 46-year-old pregnant woman had visited his clinic at the Royal Melbourne Hospital in Australia to hear the results of an amniocentesis test to screen her baby's chromosomes for abnormalities. The baby was fine — but follow-up tests had revealed something astonishing about the mother. Her body was built of cells from two individuals, probably from twin embryos that had merged in her own mother's womb. And there was more. One set of cells carried two X chromosomes, the complement that typically makes a person female; the other had an X and a Y. Halfway through her fifth decade and pregnant with her third child, the woman learned for the first time that a large part of her body was chromosomally male¹. "That's kind of science-fiction material for someone who just came in for an amniocentesis," says James.

Sex can be much more complicated than it at first seems. According to the simple scenario, the presence or absence of a Y chromosome is what counts: with it, you are male, and without it, you are female. But doctors have long known that some people straddle the boundary — their sex chromosomes say one thing, but their gonads (ovaries or testes) or sexual anatomy say another. Parents of children with these kinds of conditions — known as intersex conditions, or differences or disorders of sex development (DSDs) — often face difficult decisions about whether to bring up their child as a boy or a girl. Some researchers now say that as many as 1 person in 100 has some form of DSD².

When genetics is taken into consideration, the boundary between the

sexes becomes even blurrier. Scientists have identified many of the genes involved in the main forms of DSD, and have uncovered variations in these genes that have subtle effects on a person's anatomical or physiological sex. What's more, new technologies in DNA sequencing and cell biology are revealing that almost everyone is, to varying degrees, a patchwork of genetically distinct cells, some with a sex that might not match that of the rest of their body. Some studies even suggest that the sex of each cell drives its behaviour, through a complicated network of molecular interactions. "I think there's much greater diversity within male or female, and there is certainly an area of overlap where some people can't easily define themselves within the binary structure," says John Achermann, who studies sex development and endocrinology at University College London's Institute of Child Health.

These discoveries do not sit well in a world in which sex is still defined in binary terms. Few legal systems allow for any ambiguity in biological sex, and a person's legal rights and social status can be heavily influenced by whether their birth certificate says male or female.

"The main problem with a strong dichotomy is that there are intermediate cases that push the limits and ask us to figure out exactly where the dividing line is between males and females," says Arthur Arnold at the University of California, Los Angeles, who studies biological sex differences. "And that's often a very difficult problem, because sex can be defined a number of ways."

THE START OF SEX

That the two sexes are physically different is obvious, but at the start of life, it is not. Five weeks into development, a human embryo has the potential to form both male and female anatomy. Next to the developing kidneys, two bulges known as the gonadal ridges emerge alongside two pairs of

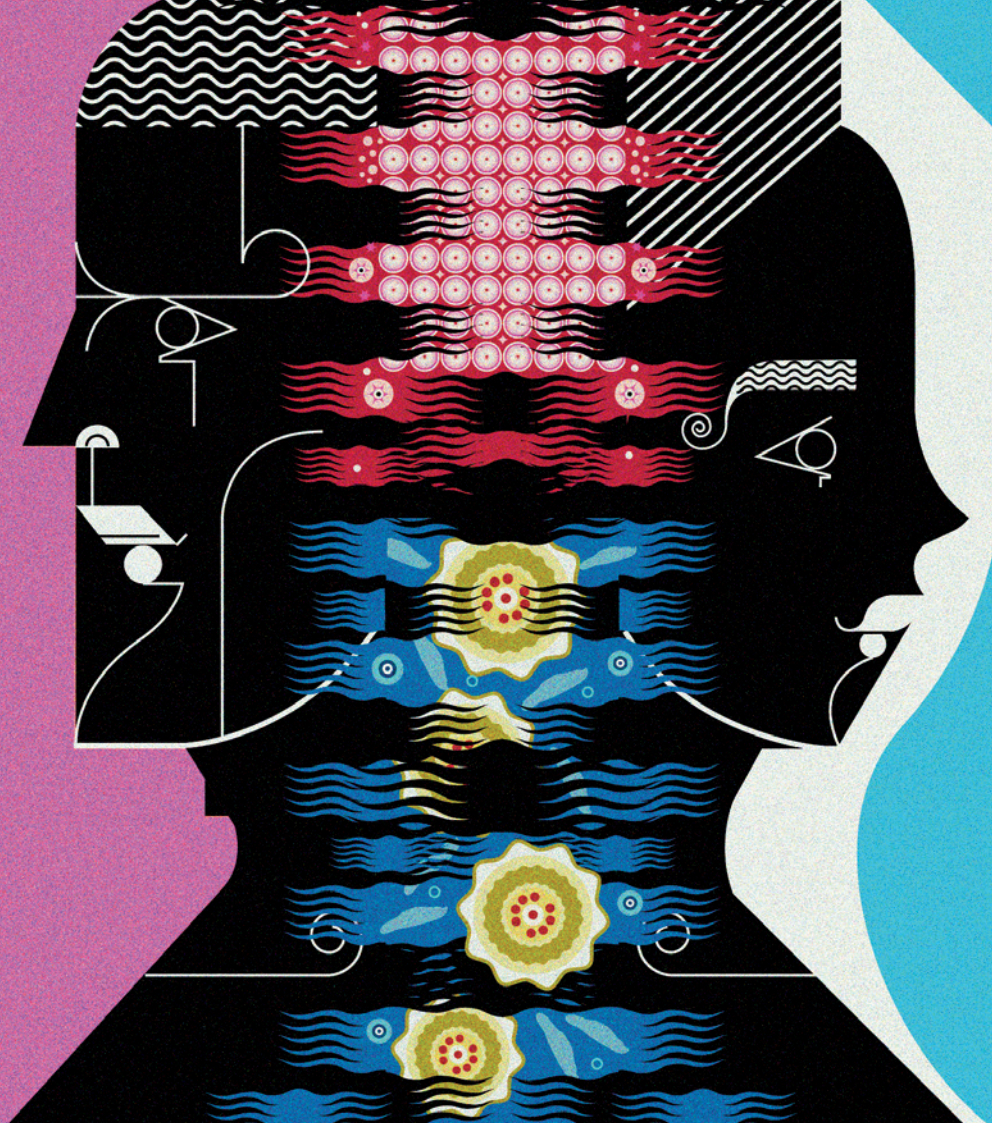


ILLUSTRATION BY JONNY WAN

ducts, one of which can form the uterus and Fallopian tubes, and the other the male internal genital plumbing: the epididymes, vas deferentia and seminal vesicles. At six weeks, the gonad switches on the developmental pathway to become an ovary or a testis. If a testis develops, it secretes testosterone, which supports the development of the male ducts. It also makes other hormones that force the presumptive uterus and Fallopian tubes to shrink away. If the gonad becomes an ovary, it makes oestrogen, and the lack of testosterone causes the male plumbing to wither. The sex hormones also dictate the development of the external genitalia, and they come into play once more at puberty, triggering the development of secondary sexual characteristics such as breasts or facial hair.

Changes to any of these processes can have dramatic effects on an individual's sex. Gene mutations affecting gonad development can result in a person with XY chromosomes developing typically female characteristics, whereas alterations in hormone signalling can cause XX individuals to develop along male lines.

For many years, scientists believed that female development was the default programme, and that male development was actively switched on by the presence of a particular gene on the Y chromosome. In 1990, researchers made headlines when they uncovered the identity of this gene^{3,4}, which they called *SRY*. Just by itself, this gene can switch the gonad from ovarian to testicular development. For example, XX individuals who carry a fragment of the Y chromosome that contains *SRY* develop as males.

By the turn of the millennium, however, the idea of femaleness being a passive default option had been toppled by the discovery of genes that actively promote ovarian development and suppress the testicular programme — such as one called *WNT4*.

NATURE.COM
For a podcast on the
sex spectrum, see:
go.nature.com/xowzq5

XY individuals with extra copies of this gene can develop atypical genitals and gonads, and a rudimentary uterus and Fallopian tubes⁵. In 2011, researchers showed⁶ that if another key ovarian gene, *RSPO1*, is not working normally, it causes XX people to develop an ovotestis — a gonad with areas of both ovarian and testicular development.

These discoveries have pointed to a complex process of sex determination, in which the identity of the gonad emerges from a contest between two opposing networks of gene activity. Changes in the activity or amounts of molecules (such as *WNT4*) in the networks can tip the balance towards or away from the sex seemingly spelled out by the chromosomes. “It has been, in a sense, a philosophical change in our way of looking at sex; that it’s a balance,” says Eric Vilain, a clinician and the director of the Center for Gender-Based Biology at the University of California, Los Angeles. “It’s more of a systems-biology view of the world of sex.”

BATTLE OF THE SEXES

According to some scientists, that balance can shift long after development is over. Studies in mice suggest that the gonad teeters between being male and female throughout life, its identity requiring constant maintenance. In 2009, researchers reported⁷ deactivating an ovarian gene called *Foxl2* in adult female mice; they found that the granulosa cells that support the development of eggs transformed into Sertoli cells, which support sperm development. Two years later, a separate team showed⁸ the opposite: that inactivating a gene called *Dmrt1* could turn adult testicular cells into ovarian ones. “That was the big shock, the fact that it was going on post-natally,” says Vincent Harley, a geneticist who studies gonad development at the MIMR-PHI Institute for Medical Research in Melbourne.

The gonad is not the only source of diversity in sex. A number of DSDs are caused by changes in the machinery that responds to hormonal

THE SEX SPECTRUM

A typical male has XY chromosomes, and a typical female has XX. But owing to genetic variation or chance events in development, some people do not fit neatly into either category. Some are classed as having differences or disorders of sex development (DSDs), in which their sex chromosomes do not match their sexual anatomy.

- Chromosomes
- Gonads
- Genitals
- Other characteristics/examples

Typical male	Subtle variations	Moderate variations	46,XY DSD
<ul style="list-style-type: none"> ● XY ● Testes ● Male internal and external genitals ● Male secondary sexual characteristics 	<ul style="list-style-type: none"> ● XY ● Testes ● Male internal and external genitals ● Subtle differences such as low sperm production. Some caused by variation in sex-development genes. 	<ul style="list-style-type: none"> ● XY ● Testes ● Male external genitals with anatomical variations such as urethral opening on underside of penis. ● Affects 1 in 250–400 births. 	<ul style="list-style-type: none"> ● XY ● Testes ● Often ambiguous ● The hormonal disorder persistent Müllerian duct syndrome results in male external genitals and testes, but also a womb and Fallopian tubes.

signals from the gonads and other glands. Complete androgen insensitivity syndrome, or CAIS, for example, arises when a person's cells are deaf to male sex hormones, usually because the receptors that respond to the hormones are not working. People with CAIS have Y chromosomes and internal testes, but their external genitalia are female, and they develop as females at puberty.

Conditions such as these meet the medical definition of DSDs, in which an individual's anatomical sex seems to be at odds with their chromosomal or gonadal sex. But they are rare — affecting about 1 in 4,500 people⁹. Some researchers now say that the definition should be widened to include subtle variations of anatomy such as mild hypospadias, in which a man's urethral opening is on the underside of his penis rather than at the tip. The most inclusive definitions point to the figure of 1 in 100 people having some form of DSD, says Vilain (see 'The sex spectrum').

But beyond this, there could be even more variation. Since the 1990s, researchers have identified more than 25 genes involved in DSDs, and next-generation DNA sequencing in the past few years has uncovered a wide range of variations in these genes that have mild effects on individuals, rather than causing DSDs. "Biologically, it's a spectrum," says Vilain.

A DSD called congenital adrenal hyperplasia (CAH), for example, causes the body to produce excessive amounts of male sex hormones; XX individuals with this condition are born with ambiguous genitalia (an enlarged clitoris and fused labia that resemble a scrotum). It is usually caused by a severe deficiency in an enzyme called 21-hydroxylase. But women carrying mutations that result in a milder deficiency develop a 'non-classical' form of CAH, which affects about 1 in 1,000 individuals; they may have male-like facial and body hair, irregular periods or fertility problems — or they might have no obvious symptoms at all. Another gene, *NR5A1*, is currently fascinating researchers because variations in it cause a wide range of effects¹⁰, from underdeveloped gonads to mild hypospadias in men, and premature menopause in women.

Many people never discover their condition unless they seek help for infertility, or discover it through some other brush with medicine. Last year, for example, surgeons reported that they had been operating on a hernia in a man, when they discovered that he had a womb¹¹. The man was 70, and had fathered four children.

CELLULAR SEX

Studies of DSDs have shown that sex is no simple dichotomy. But things become even more complex when scientists zoom in to look at individual cells. The common assumption that every cell contains the same set of genes is untrue. Some people have mosaicism: they develop from a single fertilized egg but become a patchwork of cells with different genetic make-ups. This can happen when sex chromosomes are doled out unevenly between dividing cells during early embryonic development. For example, an embryo that starts off as XY can lose a Y chromosome from a subset of its cells. If most cells end up as XY, the result

is a physically typical male, but if most cells are X, the result is a female with a condition called Turner's syndrome, which tends to result in restricted height and underdeveloped ovaries. This kind of mosaicism is rare, affecting about 1 in 15,000 people.

The effects of sex-chromosome mosaicism range from the prosaic to the extraordinary. A few cases have been documented in which a mosaic XXY embryo became a mix of two cell types — some with two X chromosomes and some with two Xs and a Y — and then split early in development¹². This results in 'identical' twins of different sexes.

There is a second way in which a person can end up with cells of different chromosomal sexes. James's patient was a chimaera: a person who develops from a mixture of two fertilized eggs, usually owing to a merger between embryonic twins in the womb. This kind of chimaerism resulting in a DSD is extremely rare, representing about 1% of all DSD cases.

Another form of chimaerism, however, is now known to be widespread. Termed microchimaerism, it happens when stem cells from a fetus cross the placenta into the mother's body, and vice versa. It was first identified in the early 1970s — but the big surprise came more than two decades later, when researchers discovered how long these crossover cells survive, even though they are foreign tissue that the body should, in theory, reject. A study in 1996 recorded women with fetal cells in their blood as many as 27 years after giving birth¹³; another found that maternal cells remain in children up to adulthood¹⁴. This type of work has further blurred the sex divide, because it means that men often carry cells from their mothers, and women who have been pregnant with a male fetus can carry a smattering of its discarded cells.

Microchimaeric cells have been found in many tissues. In 2012, for example, immunologist Lee Nelson and her team at the University of Washington in Seattle found XY cells in post-mortem samples of women's brains¹⁵. The oldest woman carrying male DNA was 94 years old. Other studies have shown that these immigrant cells are not idle; they integrate into their new environment and acquire specialized functions, including (in mice at least) forming neurons in the brain¹⁶. But what is not known is how a peppering of male cells in a female, or vice versa, affects the health or characteristics of a tissue — for example, whether it makes the tissue more susceptible to diseases more common in the opposite sex. "I think that's a great question," says Nelson, "and it is essentially entirely unaddressed." In terms of human behaviour, the consensus is that a few male microchimaeric cells in the brain seem unlikely to have a major effect on a woman.

Scientists are now finding that XX and XY cells behave in different ways, and that this can be independent of the action of sex hormones. "To tell you the truth, it's actually kind of surprising how big an effect of sex chromosomes we've been able to see," says Arnold. He and his colleagues have shown¹⁷ that the dose of X chromosomes in a mouse's body can affect its metabolism, and studies in a lab dish suggest¹⁸ that XX and XY cells behave differently on a molecular level, for example with different metabolic responses to stress. The next challenge, says

SURGEONS DISCOVERED THAT THE MAN HAD A WOMB. HE WAS 70.

	Ovotesticular DSD	46,XX testicular DSD	Moderate variations	Subtle variations	Typical female
● Chromosomes	● XX, XY or mix of both	● XX	● XX	● XX	● XX
● Gonads	● Both ovarian and testicular tissue	● Small testes	● Ovaries	● Ovaries	● Ovaries
● Genitals	● Ambiguous	● Male external genitals	● Female internal and external genitals	● Female internal and external genitals	● Female internal and external genitals
● Other characteristics/ examples	● Rare reports of predominantly XY people conceiving and bearing a healthy child.	● Usually caused by presence of male sex-determining gene <i>SRY</i> .	● Variations in sex development such as premature shutdown of ovaries. Some caused by variation in sex-development genes.	● Subtle differences such as excess male sex hormones or polycystic ovaries.	● Female secondary sexual characteristics

Arnold, is to uncover the mechanisms. His team is studying the handful of X-chromosome genes now known to be more active in females than in males. “I actually think that there are more sex differences than we know of,” says Arnold.

BEYOND THE BINARY

Biologists may have been building a more nuanced view of sex, but society has yet to catch up. True, more than half a century of activism from members of the lesbian, gay, bisexual and transgender community has softened social attitudes to sexual orientation and gender. Many societies are now comfortable with men and women crossing conventional societal boundaries in their choice of appearance, career and sexual partner. But when it comes to sex, there is still intense social pressure to conform to the binary model.

This pressure has meant that people born with clear DSDs often undergo surgery to ‘normalize’ their genitals. Such surgery is controversial because it is usually performed on babies, who are too young to consent, and risks assigning a sex at odds with the child’s ultimate gender identity — their sense of their own gender. Intersex advocacy groups have therefore argued that doctors and parents should at least wait until a child is old enough to communicate their gender identity, which typically manifests around the age of three, or old enough to decide whether they want surgery at all.

This issue was brought into focus by a lawsuit filed in South Carolina in May 2013 by the adoptive parents of a child known as MC, who was born with ovotesticular DSD, a condition that produces ambiguous genitalia and gonads with both ovarian and testicular tissue. When MC was 16 months old, doctors performed surgery to assign the child as female — but MC, who is now eight years old, went on to develop a male gender identity. Because he was in state care at the time of his treatment, the lawsuit alleged not only that the surgery constituted medical malpractice, but also that the state denied him his constitutional right to bodily integrity and his right to reproduce. Last month, a court decision prevented the federal case from going to trial, but a state case is ongoing.

“This is potentially a critically important decision for children born with intersex traits,” says Julie Greenberg, a specialist in legal issues relating to gender and sex at Thomas Jefferson School of Law in San Diego, California. The suit will hopefully encourage doctors in the United States to refrain from performing operations on infants with DSDs when there are questions about their medical necessity, she says. It could raise awareness about “the emotional and physical struggles intersex people are forced to endure because doctors wanted to ‘help’ us fit in,” says Georgiann Davis, a sociologist who studies issues surrounding intersex traits and gender at the University of Nevada, Las Vegas, who was born with CAIS.

Doctors and scientists are sympathetic to these concerns, but the MC case also makes some uneasy — because they know how much is still to be learned about the biology of sex¹⁹. They think that changing medical practice by legal ruling is not ideal, and would like to see more data collected on outcomes such as quality of life and sexual function to help decide the best course of action for people with DSDs — something that researchers are starting to do.

Diagnoses of DSDs once relied on hormone tests, anatomical

inspections and imaging, followed by painstaking tests of one gene at a time. Now, advances in genetic techniques mean that teams can analyse multiple genes at once, aiming straight for a genetic diagnosis and making the process less stressful for families. Vilain, for example, is using whole-exome sequencing — which sequences the protein-coding regions of a person’s entire genome — on XY people with DSDs. Last year, his team showed²⁰ that exome sequencing could offer a probable diagnosis in 35% of the study participants whose genetic cause had been unknown.

Vilain, Harley and Achermann say that doctors are taking an increasingly circumspect attitude to genital surgery. Children with DSDs are treated by multidisciplinary teams that aim to tailor management and support to each individual and their family, but this usually involves raising a child as male or female even if no surgery is done. Scientists and advocacy groups mostly agree on this, says Vilain: “It might be difficult for children to be raised in a gender that just does not exist out there.” In most countries, it is legally impossible to be anything but male or female.

Yet if biologists continue to show that sex is a spectrum, then society and state will have to grapple with the consequences, and work out where and how to draw the line. Many transgender and intersex activists dream of a world where a person’s sex or gender is irrelevant. Although some governments are moving in this direction, Greenberg is pessimistic about the prospects of realizing this dream — in the United States, at least. “I think to get rid of gender markers altogether or to allow a third, indeterminate marker, is going to be difficult.”

So if the law requires that a person is male or female, should that sex be assigned by anatomy, hormones, cells or chromosomes, and what should be done if they clash? “My feeling is that since there is not one biological parameter that takes over every other parameter, at the end of the day, gender identity seems to be the most reasonable parameter,” says Vilain. In other words, if you want to know whether someone is male or female, it may be best just to ask. ■

Claire Ainsworth is a freelance writer based in Hampshire, UK.

- James, P. A., Rose, K., Francis, D. & Norris, F. M. *J. Med. Genet.* **A 155**, 2484–2488 (2011).
- Arboleda, V. A., Sandberg, D. E. & Vilain, E. *Nature Rev. Endocrinol.* **10**, 603–615 (2014).
- Sinclair, A. H. *et al. Nature* **346**, 240–244 (1990).
- Berta, P. *et al. Nature* **348**, 448–450 (1990).
- Jordan, B. K. *et al. Am. J. Hum. Genet.* **68**, 1102–1109 (2001).
- Tomaselli, S. *et al. PLoS ONE* **6**, e16366 (2011).
- Uhlenhaut, N. H. *et al. Cell* **139**, 1130–1142 (2009).
- Matson, C. K. *et al. Nature* **476**, 101–104 (2011).
- Hughes, I. A., Houk, C., Ahmed, S. F., Lee, P. A. & LWPES1/ESPE2 Consensus Group. *Arch. Dis. Child.* **91**, 554–563 (2006).
- El-Khairi, R. & Achermann, J. C. *Semin. Reprod. Med.* **30**, 374–381 (2012).
- Sherwani, A. Y. *et al. Int. J. Surg. Case Rep.* **5**, 1285–1287 (2014).
- Tachon, G. *et al. Hum. Reprod.* **29**, 2814–2820 (2014).
- Bianchi, D. W., Zickwolf, G. K., Weil, G. J., Sylvester, S. & DeMaria, M. A. *Proc. Natl Acad. Sci. USA* **93**, 705–708 (1996).
- Maloney, S. *et al. J. Clin. Invest.* **104**, 41–47 (1999).
- Chan, W. F. N. *et al. PLoS ONE* **7**, e45592 (2012).
- Zeng, X. X. *et al. Stem Cells Dev.* **19**, 1819–1830 (2010).
- Link, J. C., Chen, X., Arnold, A. P. & Reue, K. *Adipocyte* **2**, 74–79 (2013).
- Penalzo, C. *et al. FASEB J.* **23**, 1869–1879 (2009).
- Warne, G. L. *Sex Dev.* **2**, 268–277 (2008).
- Baxter, R. M. *et al. J. Clin. Endocrinol. Metab.* **100**, E333–E344 (2014).



The endangered dead

The billions of specimens in natural-history museums are becoming more useful for tracking Earth's shrinking biodiversity. But the collections also face grave threats.

BY CHRISTOPHER KEMP

Ricardo Moratelli surveys several hundred dead bats — their wings neatly folded — in a room deep inside the Smithsonian Institution in Washington DC. He moves methodically among specimens arranged in ranks like a squadron of bombers on a mission. Attached to each animal's right ankle is a tag that tells Moratelli where and when the creature was collected, and by whom. Some of the tags have yellowed with age — they mark bats that were collected more than a century ago. Moratelli selects a small, compact individual with dark wings and a luxurious golden pelage. It fits easily in his cupped palm.

To the untrained eye, this specimen looks identical to the rest. But Moratelli, a post-doctoral fellow at the Smithsonian's National Museum of Natural History, has discovered that the bat in his hands is a new species. It was collected in February 1979 in an Ecuadorian forest on the western slopes of the Andes. A subadult male, it has been waiting for decades for someone such as Moratelli to recognize its uniqueness. He named it *Myotis diminutus*¹. Before Moratelli could take that step, however, he had to collect morphometric data — precise measurements of the skull and post-cranial skeleton — from other specimens. In all, he studied 3,000 other bats from 18 collections around the world.

Myotis diminutus is not alone. And neither is Ricardo Moratelli.

Across the world, natural-history collections hold thousands of species awaiting identification. In fact, researchers today find

many more novel animals and plants by sifting through decades-old specimens than they do by surveying tropical forests and remote landscapes. An estimated three-quarters of newly named mammal species are already part of a natural-history collection at the time they are identified. They sometimes sit unrecognized for a century or longer, hidden in drawers, half-forgotten in jars, misidentified, unlabelled.

"It's certainly the case that collections right now have vast resources of undescribed material," says Robert Voss, curator of mammals at the American Museum of Natural History (AMNH) in New York.

These collections are becoming increasingly valuable thanks to newly developed techniques and databases. Through DNA sequencing, digital registries and other advances, existing collections can be interrogated in new ways, revealing more about Earth's biodiversity, and how quickly it is disappearing.

But just as the collections are growing more valuable, they are falling into decline. With many institutions struggling to cope with significant budget cuts, some collections are being neglected, damaged or lost altogether. And the scientists who study them are also threatened as their positions disappear.

CUTS TO COLLECTIONS

"This is the repository of all life that we know has existed," says Michael Mares, director of the Sam Noble Museum at the University of Oklahoma in Norman, and past president of the American Society of Mammalogists. "If you want to go back and do a survey of the

mammals of Kuala Lumpur or something 30 years or 40 years ago, you can't go back," he says. "You have to go to the collections to do it."

In 1758, with the publication of the encyclopaedic *Systema Naturae*, Carl Linnaeus attempted to classify nature — an effort that continues today at almost 8,000 natural-history collections around the world. The United States alone holds an estimated 1 billion specimens, and the global figure may reach 3 billion. The average institution displays only about 1% or less of its store. The rest — often hundreds of thousands of specimens — is catalogued and stored away, inaccessible to the public.

The collections are overseen by a dwindling corps of managers and curators — mainly taxonomists who describe species, and systematists who study the relationship between organisms. The Field Museum in Chicago, Illinois, had 39 curators in 2001. Today, there are just 21. At present, there is no curator of fishes — an enormously diverse class of animal. Neither The Field Museum nor the AMNH — which hold two of the largest collections in the world — has a lepidopterist on staff, even though both collections contain hundreds of thousands of butterfly and moth specimens. Similarly, the National Museum of Natural History has seen a steady drop in the number of curators — from a high of 122 in 1993 to a low of 81 last year.

The decline is not limited to the United States. "The situation in the United Kingdom is the same, or worse," says Paolo Viscardi, chair of the UK-based Natural Sciences Collections Association and a curator at the Horniman



Ricardo Moratelli examines bat specimens in the Smithsonian Institution's National Museum of Natural History in Washington DC.

PHOTOS BY CHRIS MADDALONI/NATURE

Museum in London. Commonly, a museum will restructure its staff, replacing three or four curatorial positions with a single collections manager, and sometimes an assistant. That manager might cover every discipline, from contemporary art to the natural sciences.

Since the economic crisis of 2008, many institutions are operating with smaller budgets. The few museums that get significant numbers of research grants have shifted their science focus to molecular techniques, which are better funded than more traditional taxonomic approaches. Many museums are emphasizing education and entertainment as they cut back on curatorial staff, says Scott Schaefer, associate dean of science for collections at the AMNH. Schaefer says that he has seen significant changes in many natural-history museums since 2008. "They tend to shift away from the conduct of research to simply the telling of the story of the sciences, in the same way that Walt Disney Company may represent science as entertainment," he says.

According to Mares, most of the estimated 1,800 collections in the United States are small. "The great majority of these are hanging by a thread," he says. "They have nobody to care for them."

Even well-funded institutions are facing difficulties. At the University of Michigan in Ann Arbor, for example, one of the country's largest biodiversity collections has been warehoused in new state-of-the-art facilities, carefully maintained but difficult for researchers to access, says Voss, who did his graduate work at the university. "It's as if we decided that we

didn't want anybody doing research in our libraries anymore," he says, "but we're going to keep the books."

As curators are lost, actual specimens sometimes disappear through neglect or accidents. In 2010, a fire consumed 85,000 snake specimens and an estimated 450,000 scorpion and spider specimens at the Butantan Institute in São Paulo, Brazil.

"We see a decline in many collections in many countries," says Mares. "If a collection is sinking, no one will say it is." The concern is that administrators will get rid of collections if museum personnel point out problems, he says. "It's too dangerous. They survive by hiding."

DECADES OF WAITING

Museum staff and researchers have a name for the barriers that slow down species discovery: the taxonomic impediment. And one measure of the taxonomic impediment is the lag time — the gap between when a new species is first collected and when it is identified. Currently, the average lag time is 21 years².

It is not clear whether that lag is increasing, but it often stretches much longer than the average. In April 1856, Henry Clay Caldwell of the United States Navy found a large, fruit-eating bat on the Samoan island of Upolu. The specimen currently resides at the Academy of Natural Sciences of Drexel University in Philadelphia, Pennsylvania, and details of the find are now scarce: a few faded, hand-written descriptors on a box, a skull and a fragment of discoloured skin. In 2009, Kristofer Helgen, a mammal curator at the Smithsonian Institution, held

the skull up to the light and realized it was an unknown species. More than 150 years after it was first collected, he named the species *Pteropus allenorum* — the small Samoan flying fox³. The species is already extinct on the island.

Like Helgen, Moratelli is fascinated by natural-history collections. His interest in zoology began as a child, watching the *Wild America* documentary series on television with his father. Moratelli has described six species of bat and is preparing descriptions of eight more, all of which he found in collections. The shortest lag time was 29 years; the longest was 111.

Researchers say that such work is crucial for understanding biodiversity and how it is being threatened. "We are in the middle of a biodiversity crisis, and collections-based institutions have a unique role in society to document that biodiversity," says Quentin Wheeler, a taxonomist and president of the College of Environmental Science and Forestry at the State University of New York in Syracuse. "When we only know 10–20% of the species, we're at a huge disadvantage to detect changes in the environment, whether it's species extinctions or introductions or whatever."

The threats to museum staff and collections reflect changes that have been reshaping research for decades. With the rise of molecular biology, funding agencies and universities are providing less support for ornithologists, herpetologists, botanists and other specialized researchers who practise taxonomy. New species are still being described. But by whom?

"There are increasing numbers of non-taxonomists describing species because there are no

taxonomists doing it," says Wheeler. Instead, it has fallen to geneticists, behavioural zoologists and others not trained in taxonomy to name species. "Increasing numbers of biologists have to resort to naming them themselves or it simply won't get done."

Such careful taxonomic work is required for cataloguing biodiversity and protecting endangered species, says Helgen, who has named more than 30 species from specimens found in collections. "Every time I name one of these species," he says, "people start to think more about it; try to learn more about it; it gets on endangered-species lists."

Even with the problems facing museum collections and those who study them, there are some bright spots. The California Academy of Sciences in San Francisco is recruiting curators and growing its collection. This year, it will acquire a collection of 1.5 million weevils — a gift from a pair of scientists who wish to remain anonymous.

Museums are also trying to reach wider audiences by digitizing their collections and making them more accessible. "That's a major thrust to the Smithsonian right now," says John Kress, the institution's interim under-secretary for science. By the time the process is complete, he says, around 5 million botanical specimens — the oldest dating back to 1504 — will have been scanned. The California Academy of Sciences has partnered with Google to put images of its specimens online, along with other identifying information.

The push towards digitization will make collections more available for researchers as well as amateur taxonomists, who have described a growing number of species in recent years.

"This is the repository of all life that we know has existed."

But digitization cannot fill the role of physical collections, because not all databases include key data such as the three-dimensional scans of specimens that would allow researchers to remotely measure body parts.

Other technological advances, such as advanced DNA sequencing, are boosting the value of collections, allowing researchers to identify species that were previously indistinguishable from their closest relatives.

James Hanken, a herpetologist and director of the Harvard University Museum of Comparative Zoology in Cambridge, Massachusetts, has used DNA sequencing to study *Thorius*, a genus of pygmy salamander that is endemic to Mexico. For more than 100 years, no one was able to distinguish most *Thorius* species from each other. "They're very tiny animals," says Hanken. "They're hard to tell apart just by looking at them."

But DNA sequencing helped Hanken to describe and name 14 species, all of which have



A flood forced staff to relocate specimens at the Burpee Museum of Natural History in Rockford, Illinois.

been declared endangered by the International Union for Conservation of Nature. Usually, says Hanken, once genetic data have identified a species, he can find subtle features — in the skeleton, coloration or body size — that allow him to tell the animals apart.

In biodiversity work, researchers are increasingly using DNA barcoding, a molecular technique that relies on characteristic genetic sequences to identify a species. But a DNA barcode cannot tell a researcher anything about how a particular species of bat flies, for example.

Collections are often the best, or only, option in those cases — and that message has not been heard, either by the public or by funding agencies, say some researchers. "We're not very good at quantifying the benefits of collections," says Christopher Norris, senior collections manager at the Yale Peabody Museum of Natural History in New Haven, Connecticut. "We've not been very good historically at explaining to people in nuts-and-bolts terms why it matters that we understand biodiversity."

STORED VIRUSES

Some scientists see applications for collections beyond documenting new species and studying biodiversity. The Bernice Pauahi Bishop Museum collection in Honolulu, for example, contains millions of mosquito specimens, which might tell virologists about the dynamics of mosquito-borne pathogens. Ten years ago, says Norris, researchers assumed that preservatives would have degraded the DNA of any pathogens in a specimen. But studies are showing that it is possible to recover and analyse viral DNA from museum specimens. In 2012, researchers were able to study the evolution of a retrovirus by extracting viral DNA from 120-year-old koala skins and comparing it with DNA found in skins from the 1980s⁴.

Norris says that the same could be done

with bats to help track diseases such as Ebola. (Researchers strongly suspect that bats triggered the recent outbreak in West Africa.) "You could go into museum collections and you could prospect for viral DNA," says Norris. The AMNH alone has more than 125,000 bat specimens from around the world. "I guarantee there is something out there that is probably more scary than Ebola that we haven't encountered yet."

But thoughts of deadly diseases are far from the mind of Moratelli as he bends to his work at the Smithsonian, calipers in hand. He carefully measures another bat, enters the data into his spreadsheet and places the animal onto a tray. Measure and repeat. In cabinets within reach, he has yet more specimens on loan from museums in Pennsylvania, Louisiana and California.

Last year, while at Texas Tech University in Lubbock, Moratelli discovered what appeared to be a specimen of an unknown species of Guyanese bat. He will know for certain later this year when he travels to Canada to compare the specimen to a large collection of several hundred bats from Guyana.

A few years ago, he travelled to the French National Museum of Natural History in Paris to inspect just two specimens. In the months ahead, Moratelli will repeat the measurement process thousands of times, and he knows he will discover new species. For some of these — critically endangered bats with dwindling habitats — his findings might help to avert extinction.

For others, it is already too late. ■

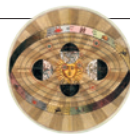
Christopher Kemp is a freelance writer in Grand Rapids, Michigan.

1. Moratelli, R. & Wilson, D. E. *Mamm. Biol.* **76**, 608–614 (2011).
2. Fontaine, B., Perrard, A. & Bouchet, P. *Curr. Biol.* **22**, R943–R944 (2012).
3. Helgen, K. M., Helgen, L. E. & Wilson, D. E. *Am. Mus. Novitates* **June**, 3646 (2009).
4. Avila-Arcos, M. C. et al. *Mol. Biol. Evol.* **30**, 299–304 (2013).

BRENT LEWIS/ROCKFORD REGISTER STAR/RRSTAR.COM

COMMENT

NEUROSCIENCE Split-brain-study pioneer writes juicy biography **p.298**



HISTORY Physicist raises hackles with reductionist approach to the past **p.300**

THEATRE Royal Shakespeare Company tackles Oppenheimer tragedy **p.301**

OBITUARY Vernon B. Mountcastle, cortical visionary, remembered **p.304**

ILLUSTRATIONS BY DAVID PARKINS



Good governance powers innovation

Corruption is a barrier to innovation, warns **Alina Mungiu-Pippidi**. Greater scrutiny of public spending is needed if science and technology are to fulfil their potential.

Former European Commission president José Manuel Barroso, in his 2013 state of Europe address, pointed to “new science studies, from new technologies” as a key to sustaining economic growth.

Similarly, US President Barack Obama stressed the importance of innovation in economic recovery in his 2014 state of the union address: “Today in America ... an entrepreneur flipped on the lights in her

tech start-up, and did her part to add to the more than 8 million new jobs our businesses have created over the past four years.” And pledges and encouragements for innovators in the developing world have come from agencies including the World Bank and World Economic Forum.

Innovation is key to prosperity. But corruption is inimical to innovation. If firms and individuals are to be creative, and if their

societies are to make the best use of that, competition and hard work must be more strongly valued than reliance on connections. My analyses show that governance that results in such societies is rarer than people think.

TOP THIRD

If you know how corrupt a country is, you can predict fairly accurately how much innovation you will see there (see ►

► ‘Virtuous circles’). In the European Union (EU), the private sector’s capacity for innovation strongly correlates with control of corruption (a correlation of 0.84), with quality of national scientific research institutions (0.85), and with gross domestic expenditure on research and development (0.9). Corruption in this analysis is defined as the abuse of public authority for private interest, resulting in a biased allocation of public resources. Control of corruption, assessed by the World Bank, is defined as the capacity of a society to restrict authorities from distributing public goods and resources in their own interests.

A country ranked below the upper third on the scale for control of corruption will not have much innovation. Of the world’s 114 democracies, only 35 are above that line, along with only 3 out of 78 countries that do not hold free elections. Romania, Bulgaria, Greece and Italy have the poorest corruption control in the EU, whereas the Nordic countries have the best, followed by the Netherlands, the United Kingdom and Germany. Outside Europe, New Zealand, Canada and Australia lead, in front of the United States, which remains the world’s most populous country where corruption is reasonably well controlled.

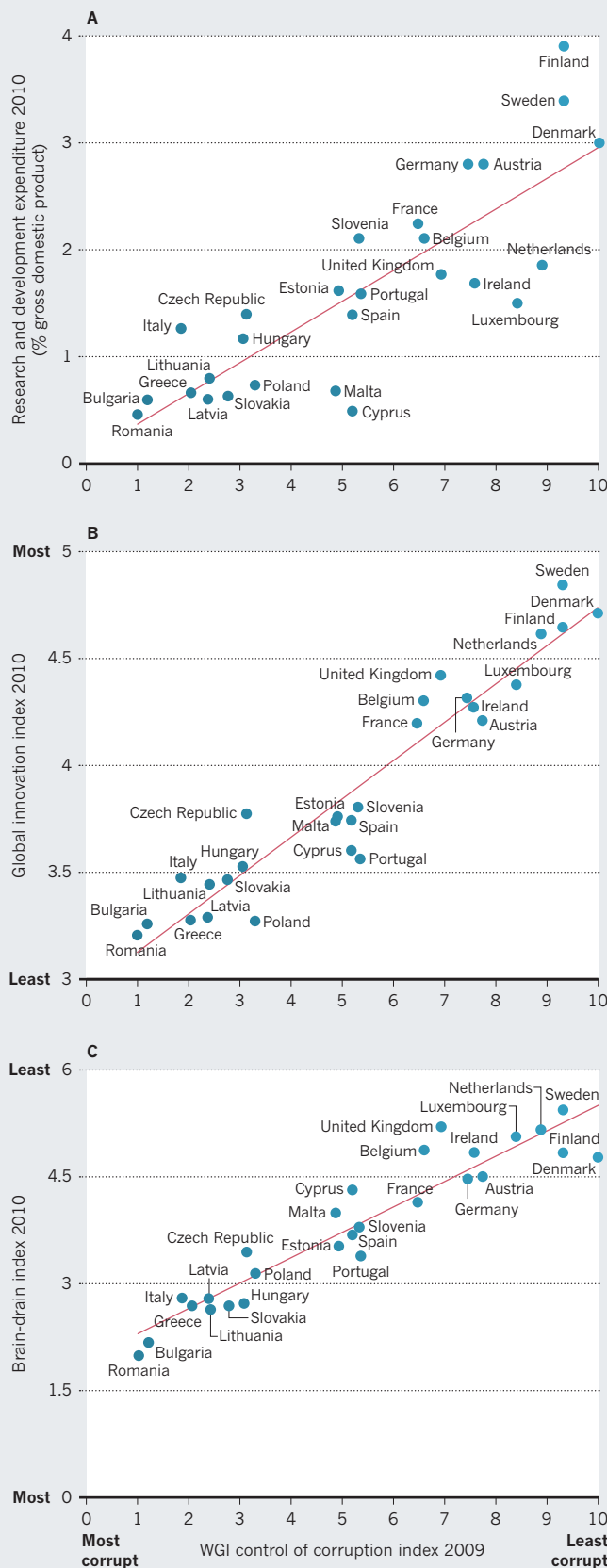
FAVOURITISM RULES

Research¹ reveals that favouritism is much more widespread than previously thought. A merit-based society takes several generations to develop, and has been achieved in only about 25 countries. These are the same nations that are on top of corruption control: those in a Nordic cluster, an Anglo-Saxon cluster, a German-speaking cluster and a few others². Only these societies have managed to create a social system in which everyone is treated similarly, so abuse of authority does not distort the allocation of public money.

Outside these 25 nations, citizens have little trust in their institutions. Nearly two-thirds of 114,000 respondents in 107 countries surveyed³ in 2013 believed that personal connections are the key factor to getting things done in the public sector — from the

VIRTUOUS CIRCLES

Research spending (A) and innovation (B) are low in European nations judged to be corrupt by the Worldwide Governance Indicators (WGI), causing talented people to flee (C). Meritocratic, democratic countries invest in science and education, which drives economic progress.



allocation of jobs in universities to the distribution of state-sponsored research funds. In the EU, a survey of roughly 85,000 respondents⁴ found that many Europeans complain of favouritism in both public and private sectors. Only in northern Europe (including France) did the majority of people believe that merit prevails. In Mediterranean countries, the two groups were nearly even; in Eastern European countries that have recently become EU members, favouritism was perceived as the dominant exchange.

Such perceptions are grounded in experiences with all aspects of school, professional and public life. Societies in which people feel this cynicism become locked in a vicious cycle: talent flees to meritocratic countries or is unproductive, further eroding their own nations’ development⁵. Although better-educated people declare themselves to be less tolerant of corruption, they offer no fewer bribes than educated ones (see go.nature.com/lmatfw).

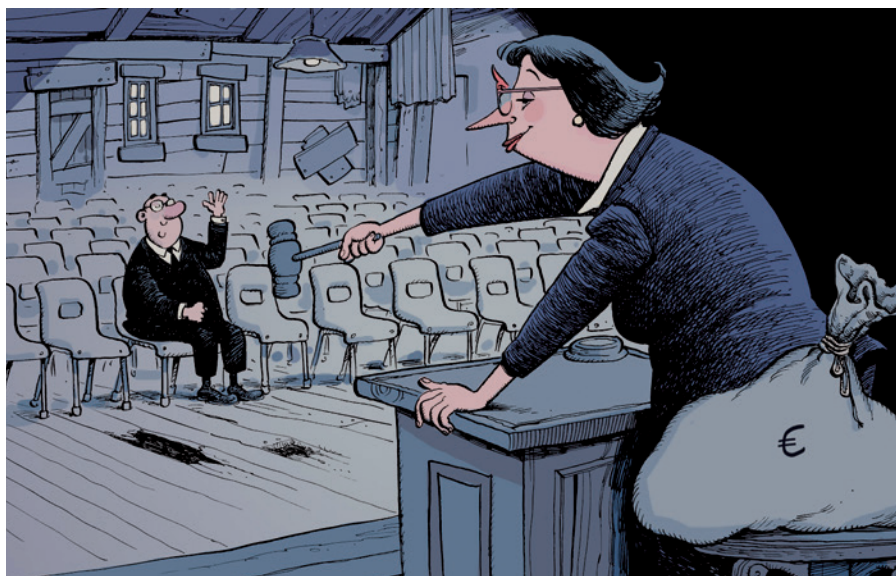
WHO WOOS WHOM

Why does innovation capacity mirror the quality of governance? Not because talent is unevenly distributed across countries and not entirely because of poverty — which, of course, plays a part because it means poorer infrastructure for innovation and technology. Romania and Bulgaria are the poorest countries in Europe, but Italy and Greece are by no means underdeveloped.

Simply, where advancement based on merit is the rule and favouritism the exception, governments and markets alike promote value, and prosperity results. In places where such a system fails to take hold, social allocation is directed preferentially rather than ethically. In these contexts, science and research are marginalized because those in power fear that talent threatens their main aim — controlling access to public and private resources.

Governments that buy political support do not invest much in education and research — the returns are seen as too general. A sports stadium or a new airport woos the companies chosen

DATA SOURCES: WORLDWIDE GOVERNANCE INDICATORS (CTRL CORRUPT); A. WORLD BANK DEVELOPMENT INDICATORS; B. GLOBAL INNOVATION INDEX; C. GLOBAL COMPETITIVENESS REPORT; ANALYSIS: A.M.-P.



CORRUPTION INDICATORS

Single bidding

During the process of bidding to acquire public funds — to win a contract for, say, building a hospital or an airport — ‘single bidding’ is when just one candidate participates and wins. In most countries, procurement legislation requires that alternative offers are sought, especially for bigger projects, so that the process is competitive and offers the best value for taxpayer money.

In corrupt countries, single bidding is common, because everyone knows that particular companies will win and others do not want to lose time or fees taking part in a sham contest. In exceptional cases, single bidding is unavoidable — NASA had no competitor in launching publicly funded space shuttles — but in most cases it is an indicator of government favouritism.

My examination of data on contract procurement from the European Union’s

Tenders Electronic Daily database (see go.nature.com/pff2nu), shows that there is much room for improvement. In Sweden, Denmark and the Netherlands, a maximum of 6% of contracts have single bidders; Croatia and Poland are at about 40%.

The research, training and education sectors raise similar concerns. In the United Kingdom, less than 3% of contracts have only one bidder in research or education projects; Poland is at 73% and 59%, respectively. Post-communist countries are doing particularly badly on this front.

Such figures are likely to be just the tip of the iceberg. How many apparently competitive tenders or postings for public-sector jobs are in fact settled behind closed doors? Countries with limited public funds must increase competitiveness for contracts before they increase the funds themselves. **A. M.-P.**

neglected. The EU needs instruments to oversee and intervene in national allocation rules.

To expose problematic practices, indicators — such as competitive distribution of research and education funds — need to be monitored closely, by pan-European watchdogs (both public and non-profit organizations) and by the European Commission. The results could be used to place conditions on a country’s participation in various European funding schemes for research, education and innovation. The European Commission has developed a detailed monitoring and advice system for member states’ economic performance, known as the European Semester. This could be expanded to include governance targets.

Much more activism is needed at the domestic level, too, from civil society, universities and local research communities. Technology helps enormously with fiscal transparency and more national civil-society watchdogs (such as those listed by the European Research Centre for Anti-Corruption and State-Building, see www.againstcorruption.eu) are needed to report on the integrity of public spending, especially on research and education. Large infrastructure projects are currently much better scrutinized than training or research grants.

For science and technology to fulfil their potential for growth, they must be empowered by a combination of funding and good governance. The impetus cannot come only from above. Reluctant national governments must be both led by the European Commission — a chief promoter of growth and innovation in Europe — and held to account by demanding domestic civil societies and science communities. ■

Alina Mungiu-Pippidi is professor of democracy studies at the Hertie School of Governance in Berlin, Germany. e-mail:pippidi@hertie-school.org

1. Mungiu-Pippidi, A. in *The Anticorruption Report*, Vol. 2: *The Anticorruption Frontline* 90–124 (ed. Mungiu-Pippidi, A.) (Barbara Budrich, 2014); available at <http://go.nature.com/toyusw>.
2. North, D. C., Wallis, J. J., & Weingast, B. R. *Violence and Social Orders: A Conceptual Framework for Interpreting Recorded Human History* (Cambridge Univ. Press, 2013).
3. Transparency International *Global Corruption Barometer 2013* (Transparency International, 2013); <http://go.nature.com/wnwjic>.
4. Charron, N. *From Åland To Ankara: European Quality Of Government Index* (The Quality of Government Institute, Gothenburg Univ., 2013); available at <http://doi.org/z78>.
5. Ariu, A. & Squicciarini, M. P. *EMBO reports* **14**, 502–504 (2013).
6. Mungiu-Pippidi, A. & Kukutschka, R. M. B. in *The Anticorruption Report*, Vol. 1: *Controlling Corruption in Europe* 14 (ed. Mungiu-Pippidi, A.) (Barbara Budrich, 2013); available at <http://go.nature.com/goo4in>.

to build it (which may contribute to the next election campaign) and the many voters who use it. A thousand science scholarships are much less profitable in these terms — they cannot be awarded to cronies with no scientific aptitude.

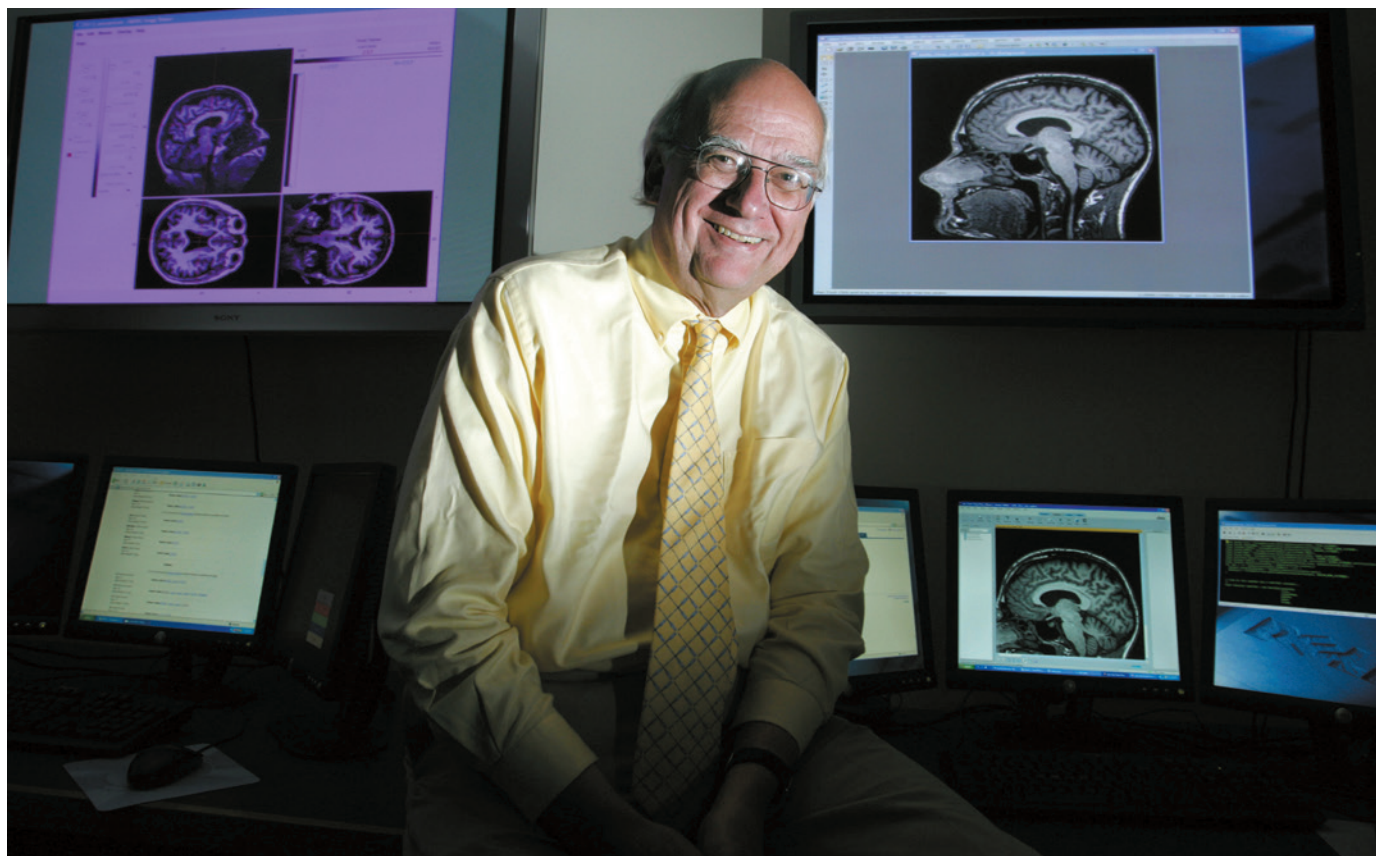
That is why more-corrupt EU states spend more on big projects such as roads and high-speed trains than on health, research, education and development (see ‘Single bidding’). When — with the best of intentions — Brussels promotes austerity policies, which funds dry up first in corrupt countries? Investment in education and science.

The European Commission knows all this, of course. Explicit recommendations

were made to EU member states not to promote austerity to areas credited in the Horizon 2020 research and innovation funding strategy with economic recovery potential. This makes it all the more concerning that Jean-Claude Juncker, the European Commission’s current president, plans to divert some of the money earmarked for research to economic stimulus. Research is what fuels development; ‘growth’ projects are mostly associated with corruption⁶.

WATCHDOGS NEEDED

The amount of public funding for research and development is frequently discussed, but the integrity of its disbursement is



RICK FRIEDMAN/CORBIS

Michael Gazzaniga pioneered research into how the brain's hemispheres can operate independently.

NEUROSCIENCE

Halving it all

Douwe Draaisma enjoys the autobiography of Michael Gazzaniga, who has studied split brains for half a century.

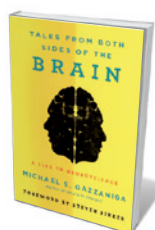
From the 1940s onwards, scores of people with intractable epilepsy were treated by surgically severing their corpus callosum, the nerve bundle that connects the left and right sides of the brain. In these 'split-brain' patients, each hemisphere operates independently. Michael Gazzaniga — known as the father of cognitive neuroscience — spent more than 50 years investigating these "splits", as he calls them affectionately in his compelling autobiography, *Tales From Both Sides of the Brain*.

As a psychology student at Dartmouth College in Hanover, New Hampshire, Gazzaniga became interested in the way brain enables mind. In the summer of 1960, he positioned himself at just the right place: Roger Sperry's lab at the California Institute of Technology (Caltech) in Pasadena. Sperry had begun a research programme on split brains, based on studies with cats and monkeys. Gazzaniga and fellow pioneer Joseph

Bogen extended this to people who had had the operation. Over the decades, as Gazzaniga relates, the programme branched out to explore perception, language, facial recognition, reasoning and many other cognitive processes. It produced a wealth of information on hemispheric specialization.

As the book unfolds, it becomes clear that split brains present a nested set of conundrums. The first is that roughly 200 million neural fibres have been cut, but nothing — apparently — happens. Memory, personality, cognition; everything is still intact.

To demonstrate that both hemispheres



Tales from Both Sides of the Brain: A Life in Neuroscience

MICHAEL S. GAZZANIGA

Ecco: 2015.

are operating separately requires shrewd experimental procedures, which Gazzaniga pioneered in the early 1960s. These revealed the second conundrum, that the left brain can see and feel things that the right brain does not, and vice versa, yet the patient experiences a single, unitary mind. Even down-right discrepancies — the right brain seeing a picture of a naked person, leaving the left brain wondering about the blush — are explained away by the mind using cleverly improvised stories.

These stories point to yet a third conundrum. Why are humans, whether with an intact or a severed callosum, so left-sided? Split-brain experiments have pointed to the existence of a 'narrator' or 'interpreter', a faculty housed in the language hemisphere (almost always the left) that explains why we behave as we do.

➔ **NATURE.COM**

For more on split brains, see:

go.nature.com/1hluvj

Unlike Bogen, who proposed some now-discredited theories on ‘left-brained’ white city dwellers and ‘right-brained’ Hopi Indians in the 1970s, Gazzaniga always kept a sober perspective on hemispheric differences. Much of his later work served to debunk the popular idea of a rational, cold-hearted left brain ranged against an emotional, intuitive right brain.

In his autobiography, Gazzaniga often seems to be a man of two minds himself. His style is colloquial and unassuming (Caltech “was chock full of mighty smart cookies and most of them could run circles around me”). He is a self-confessed big-picture man, leaving mathematics and technicalities to others. He acknowledges that the course of a career, including his own, is often steered by luck and coincidence, rather than strategy. There is also a shocking nostalgia for the days before ethical committees on animal research, when cats were gathered “from the alley”.

This cheerfully detached tone, however, is absent when Gazzaniga deals with credit and priority. His experiment with Bogen’s epilepsy patient W. J. in 1962 was the first to reveal that each hemisphere remains unaware of stimuli processed by the other. Bogen had suggested pre- and post-surgery experiments. “Thus begins a line of research that, twenty years later, almost to the day, will be awarded the Nobel Prize,” notes Gazzaniga. That 1981 prize (in Physiology or Medicine) was awarded to Sperry for his split-brain research — not to Sperry, Gazzaniga and Bogen. By then, Gazzaniga’s relationship with Sperry had become tense, and Sperry refused to let him conduct further tests on Caltech patients.

Gazzaniga writes about Sperry with much admiration and little affection. He portrays him as a fierce competitor. Gazzaniga explains that at the pioneering stage of research, ideas become inextricably mixed, and that in science — as in families — people may come away from the same event with different memories. He clearly feels that the Nobel prize should have had more than one recipient.

Gazzaniga was at the heart of a pivotal research programme and struck up friendships with neuroscience and psychology luminaries, such as David Premack, George Miller, Leon Festinger, Endel Tulving and Steven Pinker (who wrote the book’s introduction). Thus, his natural appetite to tell juicy behind-the-scenes stories is more than welcome. Historians in particular have always appreciated eighteenth-century philosopher Bernard Mandeville’s dictum that private vices can be turned to public benefit. ■

Douwe Draaisma is professor of the history of psychology at the University of Groningen in the Netherlands. His book *Forgetting will be published in March*.
e-mail: d.draaisma@rug.nl

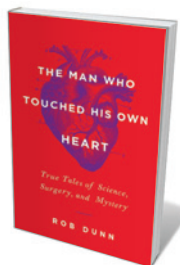
Books in brief



A New History of Life: The Radical New Discoveries about the Origins and Evolution of Life on Earth

Peter Ward and Joe Kirschvink BLOOMSBURY (2015)

Since Richard Fortey’s landmark *Life* (HarperCollins, 1997), the science on life’s origins and evolution has itself evolved. Here, palaeobiologist Peter Ward and geobiologist Joe Kirschvink weave decades of findings into an audacious retelling, hingeing on catastrophic transformation; the roles of oxygen, hydrogen sulfide and carbon dioxide as well as carbon; and the importance of ecosystems. They speculate chillingly about future impacts of the biodiversity drain, and query our own evolutionary capacity.



The Man Who Touched His Own Heart: True Tales of Science, Surgery, and Mystery

Rob Dunn LITTLE, BROWN (2015)

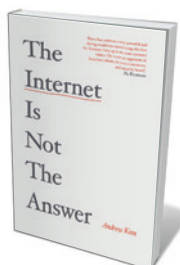
Its beat drives our lives, yet the heart — that “meat in the middle of you”, as biologist Rob Dunn puts it — remains only half understood. Dunn punctuates his chronicle of cardiac biology with stories of explorers in the “human wilderness”: nineteenth-century African American heart-surgery pioneer Daniel Hale Williams; Nobel laureate Werner Forssmann, who ran a catheter through a vein to touch his own heart; Helen Brooke Taussig, who studied avian hearts to understand human pathologies; and many more.



Is Shame Necessary?: New Uses for an Old Tool

Jennifer Jacquet PANTHEON (2015)

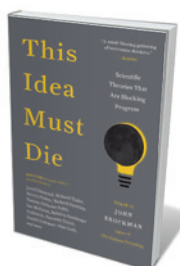
In an era when fat-cat bonuses coincide with social-service cutbacks, the baselines of shame seem to have irrevocably shifted. Yet public exposure remains a driving force for social change, argues environmental social scientist Jennifer Jacquet. In her reframing of shame, Jacquet draws on evolutionary biology, public-health research and more to examine its evolution and function, and to formulate “seven habits of highly effective shaming”. Surprises are few, but the case studies add zip — not least, the mimes hired in the 1990s by Bogotá mayor Antanas Mockus to ridicule reckless drivers.



The Internet Is Not the Answer

Andrew Keen ATLANTIC MONTHLY PRESS (2015)

Silicon Valley insider Andrew Keen joins the ranks of Internet watchers such as Nicholas Carr and Jaron Lanier with this sizzling critique of claims by the web’s supporters. Although he lauds some advances, Keen argues that industry billionaires and social-media cheerleaders create a “reality distortion field”, where wealth distribution is the rhetoric and monopolies the reality. The portraits of plutocrats running ‘disruptive’ companies in San Francisco, California — a city with 7,000 homeless people and an open-defecation problem — is a salutary reminder of the need to redefine success in a digitized world.



This Idea Must Die: Scientific Theories That Are Blocking Progress

Edited by John Brockman HARPER PERENNIAL (2015)

John Brockman, founder of virtual science salon Edge.org, gathers essays from luminaries in science and the arts for this latest in his series on the big questions of our era. This time, he asks which scientific theory is due for the dustbin. Those pitching in include neuroscientist Patricia Churchland and astronomer Martin Rees. There is plenty of pith on show, from cosmologist Max Tegmark poking holes in infinity to psychologist Paul Bloom trashing the concept of science ever maximizing happiness. **Barbara Kiser**

Unshadowed lens on the past

Robert P. Crease examines Steven Weinberg's radical retelling of the story of science.

With *To Explain the World*, Nobel-prizewinning physicist Steven Weinberg is sure to raise the hackles of professional historians of science. The book is based on lecture notes for his undergraduate courses in the history of science at the University of Texas at Austin. But he states at the outset: "I am a physicist, not a historian". He is unapologetic about judging past science from the viewpoint of the present, and scornful of scholars who view scientific results as historical or cultural products. He focuses almost exclusively on Western science (including that of medieval Islam). Although other civilizations generated much scientific knowledge, Weinberg explains, the scientific method — a special technique that "allows us to learn reliable things about the world" — was discovered and exploited first in the West.

The result is unique and provocative: imagine a history of architecture that judged edifices by the extent to which they met modern needs and building codes. Weinberg demotes many luminaries in the pantheon of science history, including philosopher René Descartes and early empiricist Francis Bacon. He elevates others, such as Aristarchus of Samos — the classical proponent of heliocentrism — and early-modern chemist Robert Boyle, an exponent of the "new aggressive style of experimental physics".

A strength of the book is its knowledgeable assessments of mechanical and astronomical systems, including those of Nicolaus Copernicus and Isaac Newton. Included is a valuable, 100-page-long set of technical notes covering the mechanical, optical and astronomical issues of early science, such as derivations of the law of refraction and the mathematics of planetary orbits.

Weinberg can be a wise and witty writer, as shown by his popular classic on the origins of the Universe, *The First Three Minutes* (Basic, 1977). In *To Explain the World*, he discusses the rejection of Aristotelian science at the University of Paris in the thirteenth century thus: "the condemnation saved science from dogmatic Aristotelianism, while the lifting of the condemnation saved science from dogmatic Christianity". He often derives lessons about science from history, illustrating them with twentieth-century examples. He notes that Copernicus's work shows "that a simple and beautiful theory that agrees pretty well with observation is often closer to the truth than a

complicated ugly theory that agrees better with observation". He then tells a story from the history of quantum mechanics, involving Erwin Schrödinger's method for calculating the energy states of hydrogen atoms. After charging the eleventh-century Persian scholar al-Biruni with using misplaced precision in calculating Earth's radius, Weinberg describes an episode when, as a summer intern, he calculated magnetic-field measurements to eight meaningless significant figures.

The approach has weaknesses. Bacon and Descartes did often err in scientific judgement, but they defended science in its infancy and helped to establish it as an intelligible and useful activity, creating a cultural niche for Weinberg's profession. These achievements are negligible only in a very narrowly conceived history of science.

Weinberg admits that he feels more at home with physics from the seventeenth century on, after the scientific method was established. His discomfort with earlier periods shows when he sometimes carelessly fails to appreciate the context of a figure or a statement. Most strikingly, he claims that Socrates was "not very interested in natural science". He bases this on a passage in Plato's *Phaedo* in which the philosopher

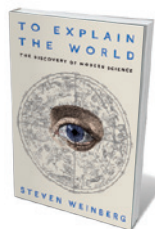
expresses disappointment with his predecessor Anaxagoras's description of heavenly bodies "in purely physical terms, without regard to what is best". But there is more to that story. In the *Phaedo*, Socrates notes that he had once embraced Anaxagoras's view of a Universe ruled by a divine mind, but later rejected it because it failed to show how or why the mechanics of the cosmos are the inevitable choices of that mind — which is in any case ultimately unknowable. So he developed his own method of investigation, starting with the strongest-looking hypothesis and testing it through questioning.

Weinberg thus cites a remark from a position that Socrates explicitly states that he has abandoned. And although Socrates' key terms, such as hypothesis and logic, do not mean for us exactly what they did for the Greeks, the philosopher was putting into motion something that Weinberg fails to recognize. His open-ended, hypothetical method of inquiry rejected foundation in pure reason or divine knowledge and, combined with mathematics, is seen by many scholars as an early formulation of the scientific method.

Weinberg opens *To Explain the World* with an excerpt from John Donne's poem 'A Lecture Upon the Shadow'. Two lovers are talking in the morning; gradually their shadows shorten, then finally vanish as the Sun moves directly overhead. The last line that Weinberg quotes is: "to brave clearness all things are reduced". He concludes his book with unapologetic praise of reductionism as the right path for science, providing "a view of why the world is the way it is".

The reductionist approach clarifies many bits and pieces of the past, such as key aspects of early astronomical and optical models. Weinberg also displays a much deeper and more intuitive insight into scientific practice than many historians and philosophers. "We learn how to do science," he writes, "not by making rules about how to do science, but from the experience of doing science, driven by desire for the pleasure we get when our methods succeed in explaining something". Still, sometimes you have to see the shadows — how something fits into its surroundings — to see it as it is. ■

Robert P. Crease is a professor of philosophy at Stony Brook University in New York, and co-author of *The Quantum Moment*.
e-mail: robert.crease@stonybrook.edu



To Explain the World: The Discovery of Modern Science
STEVEN WEINBERG
Allen Lane/Harper
Collins: 2015.



Nicolaus Copernicus's heliocentric system, illustrated in 1660.



John Heffernan (foreground) takes the leading role in *Oppenheimer*.

THEATRE

Atomic tragedian

Philip Ball sees something of Macbeth in a play about J. Robert Oppenheimer, leader of the Manhattan Project.

He rose to prominence as head of the epochal Manhattan Project, and fell as a suspected communist sympathizer during the McCarthyite 1950s: the trajectory of J. Robert Oppenheimer's life resembled that of a Shakespearean tragic hero. That was playwright Tom Morton-Smith's pitch to the Royal Shakespeare Company three years ago. Morton-Smith's new play, *Oppenheimer*, fulfils that promise: its protagonist has more than a touch of Macbeth about him.

Oppenheimer emerges as a man driven by boundless ambition — and sometimes encouraged by his wife — to compromise his integrity. He ends haunted with guilt, both for betraying friends, colleagues and lovers, and for masterminding the most destructive weapons ever made. Macbeth's "I am in blood/Stepp'd in so far" is almost too apt. And yet like Macbeth, he is not really a villain, but a character who remains sympathetic even as his flaws are exposed. For its complex portrayal of the dilemmas and ambiguities faced by the early nuclear scientists, *Oppenheimer* deserves much praise.

The play covers just the period from the discovery of nuclear fission in 1938 to the bombings of Hiroshima and Nagasaki in 1945. The 1954 hearings at which Oppenheimer's security clearance was withdrawn — largely because of a youthful flirtation with

leftist politics — can only be foreshadowed. The contrast with Heinar Kipphardt's 1964 play *In the Matter of J. Robert Oppenheimer* is interesting. Written when public concerns about atomic scientists' morality were at their peak, that work echoed Friedrich Dürrenmatt's 1961 play *The Physicists* in implying that these researchers might have been "traitors to the spirit of science". *Oppenheimer* presents a more nuanced view, giving a sense both of the diverse attitudes of Manhattan Project scientists and of the tensions between the researchers and their military leaders. Some wrestle with their consciences; others focus only on the science.

To get across the basics of nuclear fission and fusion, Morton-Smith has to resort to some highly theatrical blackboard lectures. It is neatly done, but demands that the stage scientists speak in ways that real ones never do: they become an orchestrated chorus of excited pedagogical voices. This is a common problem for science plays, and is perhaps best solved by weaving the concepts into the narrative in the 'show, don't tell' style used by Michael Frayn to demonstrate quantum

Oppenheimer
TOM MORTON-SMITH
The Swan Theatre,
Stratford-upon-Avon,
UK
Until 7 March.

► **NATURE.COM**
For more on science
in culture, see:
[nature.com/
booksandarts](http://nature.com/booksandarts)

uncertainty in *Copenhagen* (1998), or by Tom Stoppard to convey chaos theory in *Arcadia* (1993). In my view, a nuclear-physics primer is not essential here anyway — we need to know only that the researchers at Los Alamos in New Mexico are making a bomb of awesome destructive potential, and that it is hard.

The scientists themselves are captured more satisfyingly: the decent, gently witty Hans Bethe (Tom McCall), the principled Robert Wilson (Jack Holden). Morton-Smith writes Edward Teller (Ben Allen) as an almost comically monstrous egotist who considers the messy engineering of the Little Boy and Fat Man fission bombs beneath him and is determined to press ahead with the cleverer science of a thermonuclear hydrogen bomb. Walk-on parts for the more famous or infamous names — Albert Einstein, Richard Feynman, Klaus Fuchs — are a little gratuitous, but that is a quibble.

As *Oppenheimer* himself, John Heffernan captures the man's charisma and icy solipsism in a subtle and compelling performance. Creating a character who has iron in his soul, yet who inspires great devotion, is no mean feat.

Oppenheimer's brilliance was never in doubt — Bethe said that he "did more than any other man to make American theoretical physics great". Yet genius does not always ensure good life choices. Even by his own admission, Oppenheimer's bungling subterfuge in a 1943 incident in which a colleague at the University of California, Berkeley, floated the idea of getting technical information to the Soviets — a key element in the 1954 trial — was "idiotic". Oppenheimer thought deeply about the role of science in society, yet his response was to retreat into grand generalities about science's amorality: "In most scientific study, questions of good and evil, or right and wrong, play at most a minor and secondary part ... The true responsibility of a scientist ... is to the integrity and vigor of his science."

It is precisely these contradictions that make Oppenheimer ripe for theatrical exploration. His is the type that so often rises to the top in times of conflict. In his strengths and failings there are parallels with a very different man: Winston Churchill (see R. Rhodes *Nature* **501**, 488–490; 2013). "Los Alamos might have succeeded without him," Bethe wrote, "but certainly only with much greater strain, less enthusiasm, and less speed ... He brought out the best in all of us." Perhaps only someone with the charm and intellect, sense of superiority and assurance, and armour-plated flaws could have done what Oppenheimer did. He really did change the world, but it is for the rest of us to work out what to do about it. ■

Philip Ball is a writer based in London, and author of *Serving the Reich*.
e-mail: p.ball@btinternet.com

Correspondence

Early antibiotic from a cranberry bog

Losee Ling and colleagues refer to Selman Waksman's "platform of natural product drug discovery" (*Nature* **517**, 455–459; 2015; see also K. Lewis *Nature* **485**, 439–440; 2012), which alludes to Waksman's 1943 discovery from soil of streptomycin, the first drug effective against tuberculosis. In fact, René Dubos, Waksman's former student, had isolated the first antibiotic from soil bacteria more than a decade earlier.

In 1930, Dubos isolated an enzyme from an unnamed bacillus found in an acidic bog in New Jersey in which cranberry plants were growing. This enzyme destroyed the polysaccharide wall of type III *Streptococcus pneumoniae*, enabling it to both cure and protect animals infected with this streptococcus (C. L. Moberg *René Dubos, Friend of the Good Earth*; ASM Press, 2005).

Dubos went on to extract the antibiotics tyrothricin and gramicidin from the soil bacterium *Bacillus brevis* in 1939. These drugs were produced commercially and used clinically in 1940, before penicillin became available.

Waksman used Dubos' soil-enrichment technique to isolate streptomycin. He later acknowledged Dubos' discovery of gramicidin as "the stimulus which flooded with bright light the whole previously unilluminated field of the study and application of antibiotics" (S. A. Waksman in *Frontiers in Medicine* 99–119; Columbia Univ. Press, 1951). **Carol L. Moberg** *The Rockefeller University, New York, USA.* moberg@rockefeller.edu

Rescue Eastern Europe's collections

The political collapse of Eastern Europe has ravaged its priceless natural history collections. The European Union could rescue many of these as part

of its commitment to preserve cultural heritage.

In the small, war-torn nations of the western Balkans, for example, natural history collections receive much less government funding than museums of history, ethnography and archaeology. Ministries of science in countries of the former Yugoslavia dismiss the importance of natural history collections for research infrastructure or as scientific heritage.

Unstable funding forced the Sarajevo museum to close in 2012, and many of its historical specimens — including 10,000 bird skins and 500,000 insects from the Balkans — have not been properly curated for years. And the break-up of the Soviet Union left individual states with little interest in maintaining their collections — including one of the world's most important, at the Zoological Museum in St Petersburg, Russia.

All of these once well-managed collections are now decaying, making the plight of Italy's natural history museums the tip of an iceberg (see *Nature* **515**, 311–312; 2014).

Boris Kryštufek *Slovenian Museum of Natural History, Ljubljana, Slovenia.*

Nataliya Abramson *Zoological Institute, Russian Academy of Sciences, St Petersburg, Russia.*

Dražen Kotrošan *National Museum, Sarajevo, Bosnia and Herzegovina.* bkrystufek@pms-lj.si

Tweak Chinese law to end ivory demand

The proposed revisions to China's wildlife protection law of 1988 should aim to make the private ownership of threatened species illegal. This would help to control the country's flourishing trade in illegal animal products. It is also important to tackle the public's demand for such goods by changing their perceived 'luxury' status.

One of us (Z.-M.Z.) investigated criminal trading in China of ivory, rhino horn and the teeth, bones and pelts of big cats in 2010–13, all products from animals of crucial conservation concern. However, these are much sought after and have a high market value — accounting for roughly half the worth of all 78 species investigated.

Farming or ranching of these animals might present one solution, but could also act as a cover for illegal trading. Better education of the public is needed to drive home the shocking consequences of this trade for biodiversity, and to remove the demand. Coupled with amendments to the wildlife protection law, China would then be taking a major step towards valuing healthy ecosystems above *objets d'art* and traditional medicine practices. **Zhao-Min Zhou*** *Yunnan Public Security Bureau for Forests, Kunming, China.*

zhouzm81@gmail.com

*On behalf of 6 correspondents (see go.nature.com/y9f8ui for full list).

'Simple' or 'elegant' criteria are not valid

Counter to the impression given by George Ellis and Joe Silk, I have never used or endorsed the slogans "elegance will suffice" and "post-empirical science" regarding theories in fundamental physics (*Nature* **516**, 321–323, 2014). In fact, both contradict my position.

I do not think that criteria such as simplicity or elegance provide a workable basis for judging a theory's chances of viability. I seek arguments that are more reliable.

Nor is my concept of non-empirical theory confirmation driven by a wish to declare empirical data obsolete. Rather, it aims to account for the actual situation in modern fundamental physics by extending the concept of theory

confirmation while preserving the primacy of empirical data.

A theory's non-empirical confirmation relies on experimental confirmation in three ways. First, a theory's viability is defined in terms of its empirical confirmation. Second, non-empirical confirmation will always remain weaker than conclusive empirical confirmation. And third, non-empirical confirmation relies on the observation that related theories in the field were empirically confirmed.

Terminating empirical confirmation in a research field would thus eventually destroy the basis for non-empirical confirmation as well.

Richard Dawid *Ludwig Maximilian University of Munich, Germany.* richard.dawid@univie.ac.at

Czech centre marks Mendel anniversary

This month marks the 150th anniversary of Gregor Mendel's presentation of his famous study 'Experiments in plant hybrids' at a meeting of the Nature Research Society in Brno, Moravia (in today's Czech Republic). His lecture was published a year later in the society's journal.

Often called the father of modern genetics, Mendel and his scientific and cultural legacy are being honoured at the Mendelianum Centre of the Moravian Museum, the original premises of the society where his ideas were first formulated and discussed. The centre, which will be officially opened on 8 March, is both a Mendel museum and an outreach venue for modern genetics research, science and education (see www.mendelianum.cz).

Anna Matalová, Eva Matalová *Mendelianum Centre of the Moravian Museum; and Institute of Animal Physiology and Genetics, Brno, Czech Republic.* matalova@iach.cz

Vernon B. Mountcastle

(1918–2015)

Discoverer of the repeating organization of neurons in the mammalian cortex.

Vernon Benjamin Mountcastle pioneered the study of the physiological properties of single neurons in anaesthetized and awake animals. In so doing, he discovered the columnar organization of a part of the mammalian brain known as the neocortex. His groundbreaking investigations of the neural mechanisms of attention and action led to our modern understanding of how information is represented and processed in the cortex.

Mountcastle, who died on 11 January, was born in 1918 in Shelbyville, Kentucky. When he was three years old his family moved to Virginia, which was always 'home' for him. He was proud to trace his ancestry back to Pocahontas, daughter of the Native American chief Powhatan. His paternal grandfather fought in the American Civil War and survived an operation conducted by his own brothers to remove a bullet. Following in this family tradition, Mountcastle trained at Johns Hopkins School of Medicine in Baltimore, Maryland. After graduating in 1942, Mountcastle served as a physician in the US Navy Amphibious Forces during the Second World War.

He had two diplomatic successes after demobilizing in 1946: a schoolteacher named Nancy Clayton Pierpont agreed to marry him, and the neurophysiologist Philip Bard agreed to take him on as a researcher in the department of physiology at Hopkins, despite his never having done an experiment. Mountcastle was motivated by what he called "the expectation of excellence" — others' assumption that he was a better investigator than he thought he was. In the collegial environment of Hopkins, he flourished.

Mountcastle wanted to understand how the information from the sensory receptors in the skin and joints is represented in the mammalian brain. He used a microelectrode to record the responses of one neuron after another in the grey matter of the primary sensory cortex in anaesthetized cats. He discovered that any given neuron responded to only one of three types of stimulus: light touch, pressure or joint movement. These different functional types of neuron, he found, were segregated in a vertical organization, which he called a cortical 'column'. This high degree of order was unexpected. It led Mountcastle to develop the now widely accepted idea that the neocortex is built of repeated units



of the same 'canonical' local circuit — the elementary unit of cortical function.

He published a landmark paper describing these discoveries in the *Journal of Neurophysiology* in 1957. This was quickly followed in 1959 by four papers co-authored with Tom Powell, an anatomist at the University of Oxford, UK, showing that the macaque monkey cortex had a similar organization. Mountcastle's discoveries alerted his neighbours at Hopkins, David Hubel and Torsten Wiesel, to look for columns in the primary visual cortices of cats and monkeys, which they found a few years later.

Once Mountcastle had established the existence of cortical columns, the stage was set for him to link the properties of single neurons to perception. He knew that tactile sensation is more acute when a finger is moved over a surface than when it is held motionless against one. Throughout the 1960s, he and his many postdocs explored how the transient and sustained signals sent from the sensory nerves of the hand are processed in human and monkey brains. Mountcastle concluded that our experience of touch is determined mostly by the properties of the peripheral nerves, which are relayed with surprisingly high fidelity to the primary sensory cortex.

In the early 1970s, with the help of neuroscientist Edward Evarts at the US National Institute of Mental Health in

Bethesda, Maryland, Mountcastle learned to record from awake monkeys. During one experiment on the primary sensory cortex, he became so exasperated when the monkey's focus of attention seemed to have no influence on a neuron's activity that he spontaneously moved the microelectrode to a nearby region known as the parietal cortex. Suddenly he saw what he had been hoping for — the activity of the neurons depended on whether the monkey attended to the stimulus.

Mountcastle's fortuitous discovery sparked a new field of research on cortical mechanisms of directed attention, spatial perception and action. It also gave him insights into the syndrome in which damage to the right parietal lobe alters a person's perception of space and leads them to neglect the left side of their body and the surrounding space.

Vernon's commitment to his experiments remained undiminished throughout his career — even after becoming director of the Hopkins physiology department in 1964 and accumulating numerous chief editorships for, among others, the *Journal of Neurophysiology* and the textbook *Medical Physiology*. In 1969, he was elected president of the newly formed Society for Neuroscience.

Vernon said that with the help of his administrator, Mary Hilda Counselman, he could clear his desk by 9 a.m. and head for the laboratory. He regarded any neuroscientist who worked less than 60 hours a week as a 'part-timer', but as an old-school physiologist, he did not leave his assistants to do his experiments for him. Fledgling scientists found him "fierce — but very kind" and experienced the same "strong pull from above" that he said he had felt from his own mentors.

A reviewer of one of Vernon's books once opined that he was "a narrow-minded master". Eric Kandel, a Nobel prizewinner in medicine, replied aptly to such criticism of Vernon's monumental contribution: "You are right, it does not apply to the kidney or the spleen ... It only helps to explain the workings of the mind." ■

Kevan Martin is co-director of the Institute of Neuroinformatics at the University of Zurich and the Swiss Federal Institute of Technology Zurich, Switzerland.
e-mail: kevan@ini.uzh.ch

The cortical connection

Neurons in the brain's visual cortex receive inputs from thousands of other neurons. But it now emerges that each is strongly connected to only a few others: those most similar to itself. [SEE LETTER P.399](#)

BENJAMIN SCHOLL & NICHOLAS J. PRIEBE

Social-networking websites such as Facebook allow us to connect with people all over the world, but as our list of friends grows, it becomes difficult to read and react to posts from everyone. Instead, we often choose to respond primarily to those people with whom we share common interests, perhaps close family and friends. Similarly, neurons in the brain's cerebral cortex receive between 1,000 and 10,000 synaptic connections, from hundreds or thousands of other neurons¹. But it is not known whether neurons listen to all of those inputs equally. Could certain inputs be more influential than others? In this issue, Cossell *et al.*² (page 399) demonstrate that neurons in the visual cortex listen to a subset of their synaptic inputs, and that the neurons comprising that subset have strikingly similar functional properties to the target neuron — much like how our closest Facebook friends have similar interests to our own.

For decades, neuroscientists have attempted to uncover the relationship between a neuron's functional properties and its synaptic input from other neurons. On the one hand, studying this relationship *in vivo* is challenging, because although it is possible to determine which sensory stimuli excite each neuron (known as its receptive field), it is difficult to identify synaptic inputs. On the other hand, the gold standard for measuring synaptic inputs is to make direct measurements *in vitro* using simultaneous recordings from pairs of neurons in brain slices, where it is not possible to measure the receptive field. Cossell and colleagues confront this challenge by combining measurements taken *in vivo* and *in vitro* to reveal the receptive fields of synaptically connected neurons. The authors discover that neurons receive inputs from a diverse population of neighbours, but that most inputs are drowned out by a few dominant synapses.

Given the enormous number of synaptic connections identified from anatomical measurements, it is puzzling that only a few synaptic inputs dominate. Cossell and co-workers uncover a direct relationship between connection strength (known as a synaptic weight) and receptive field in connected pairs of neurons.

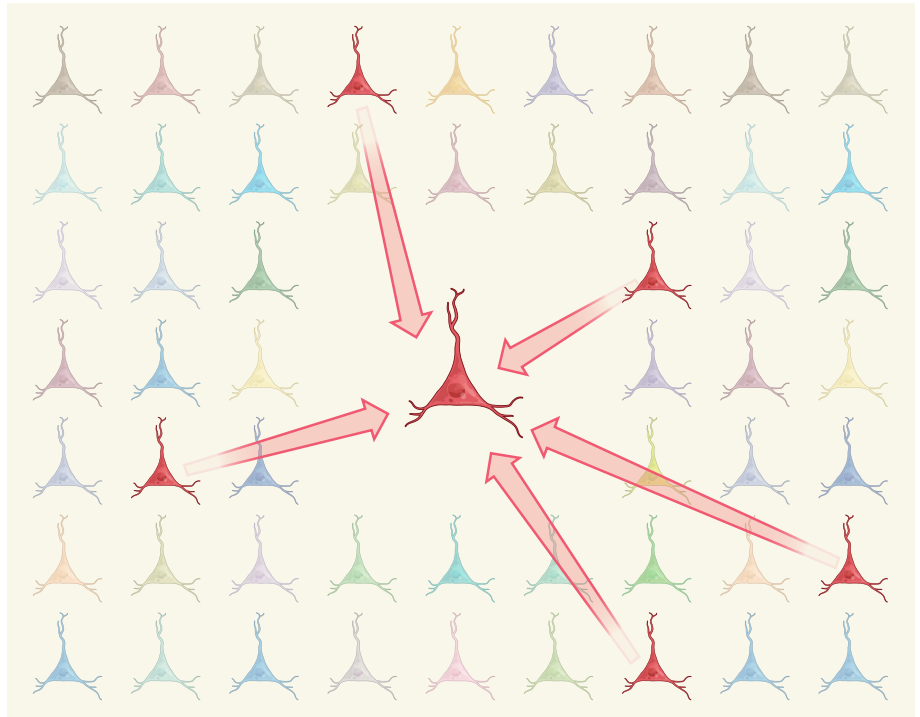


Figure 1 | Deciphering neuronal networks. Neurons in the visual cortex receive synaptic inputs from many other neurons, here represented in a grid. Neurons activated by similar sensory stimuli — that is, those with similar receptive fields — are depicted in the same colour. The large target neuron in the centre receives connections from all the other neurons in the grid (connections not shown). If all of those connections are weighted equally, it is difficult to determine the function of the target neuron. But Cossell *et al.*² report that the target neuron receives strong inputs (indicated by arrows) only from neurons with similar receptive fields to its own.

The few strong synaptic connections originate from neurons with receptive fields matching the neuron onto which they synapse. By contrast, many — but weak — synaptic connections originate from presynaptic neurons with different receptive fields (Fig. 1).

If only a few synaptic inputs drive neuronal responses, why is there such dense connectivity in the cerebral cortex? Perhaps this wiring is useful for brain development and plasticity (changes in neuronal connections), or even for the formation of memories. The presence of many weak connections might allow cortical networks to be reprogrammed easily. For instance, modulating the weighting of just a few key synaptic inputs could induce dramatic changes in a neuron's receptive field.

Changing the receptive field of cortical neurons by generating new anatomical

connections would take time and resources. By having those connections in place, and to some degree dormant, neuronal plasticity can be achieved quickly. Furthermore, because the inputs are already in place, changes in synaptic weight during associative learning can be driven by mechanisms such as Hebbian plasticity, in which correlated activity of a presynaptic neuron with that of the postsynaptic neuron increases the synaptic weight^{3,4}. An excess of synaptic contacts might therefore be key to plasticity, which occurs in mature cortical circuits^{5,6}.

Both the potential plasticity of cortical circuits and the non-uniform nature of synaptic input strengths present challenges to connectome projects, which aim to determine the functional properties of neurons on the basis of anatomical connectivity alone^{7–9}. Consider

an attempt to decode the personality traits of a single person on the basis of their full set of Facebook friends. This would be a tough job, given the wide variety of individuals in the group, which might include friends, family and acquaintances. Similarly, if we assume that all connections are equivalent in a neuronal network, predicting the response selectivity of a cortical neuron may be difficult, because the inputs are so diverse. Although connectome projects will certainly generate valuable statistics about connectivity in the cerebral cortex, Cossell and colleagues demonstrate that identifying the weights of synaptic connections is essential to account for neuronal responses. When connectivity is considered in conjunction with synaptic weighting, it should be possible to predict response selectivity.

Despite the difficulties associated with

determining connection weight from anatomical measurements alone, there are hints that not all synaptic inputs are physically equal. Properties of synaptic contacts that may be linked to synaptic weight¹⁰ include the size of postsynaptic structures such as dendritic spines, the number of neurotransmitter molecules available for release by the presynaptic terminal, and the extent to which subcellular compartments that are responsible for protein synthesis can form new postsynaptic structures. Understanding these links will be crucial for bridging the gap between functional and anatomical connectivity, so that neuroscientists can get closer to obtaining a functional connectome. ■

Benjamin Scholl and Nicholas J. Priebe
are in the Department of Neuroscience,

The University of Texas, Austin,
Texas 78712, USA.
e-mail: nico@austin.utexas.edu

1. Binzegger, T., Douglas, R. J. & Martin, K. A. C. *J. Neurosci.* **24**, 8441–8453 (2004).
2. Cossell, L. *et al. Nature* **518**, 399–403 (2015).
3. Isaac, J. T. R., Nicoll, R. A. & Malenka, R. C. *Neuron* **15**, 427–434 (1995).
4. Liao, D., Hessler, N. A. & Malinow, R. *Nature* **375**, 400–404 (1995).
5. Allard, T., Clark, S. A., Jenkins, W. M. & Merzenich, M. M. *J. Neurophysiol.* **66**, 1048–1058 (1991).
6. Sato, M. & Stryker, M. P. *J. Neurosci.* **28**, 10278–10286 (2008).
7. Kleinfeld, D. *et al. J. Neurosci.* **31**, 16125–16138 (2011).
8. Lichtman, J. W., Livet, J. & Sanes, J. R. *Nature Rev. Neurosci.* **9**, 417–422 (2008).
9. Oh, S. W. *et al. Nature* **508**, 207–214 (2014).
10. Bourne, J. N. & Harris, K. M. *Annu. Rev. Neurosci.* **31**, 47–67 (2008).

This article was published online on 4 February 2015.

ASTROPHYSICS

A lithium-rich stellar explosion

The contribution of explosions known as novae to the lithium content of the Milky Way is uncertain. Radioactive beryllium, which transforms into lithium, has been detected for the first time in one such explosion. SEE LETTER P.381

MARGARITA HERNANZ

The origin of lithium observed in today's Universe is a long-standing problem. It is known that a fraction of this light chemical element was created during the Big Bang, along with hydrogen and helium, and that another fraction has formed since then through nuclear reactions induced by energetic cosmic rays. But comparison of chemical-evolution models and observed stellar lithium abundances in the Milky Way indicates that part of the lithium should also have been synthesized in old low-mass stars, such as red giants, and in stellar explosions known as novae. However, although lithium has been observed in giants, its detection in novae has remained elusive. On page 381 of this issue, Tajitsu *et al.*¹ provide the first observational evidence of lithium synthesis in novae. The authors detected radioactive beryllium-7 (⁷Be), the parent nucleus of lithium-7 (⁷Li), during a nova explosion called V339 Del (Nova Delphini 2013).

It has long been known that almost all of the chemical elements are produced in stars by the nuclear fusion of light elements into heavier ones, starting with hydrogen fusion². The synthesized elements can then be expelled to the interstellar medium — from which new stars will form — either by stellar winds or

during supernova explosions and their dimmer relatives, novae. However, the main origin of the light elements lithium, beryllium and boron is not linked to nuclear reactions in stars. Instead, it is related to nucleosynthesis processes that are less efficient than stellar ones. This is why these elements are much

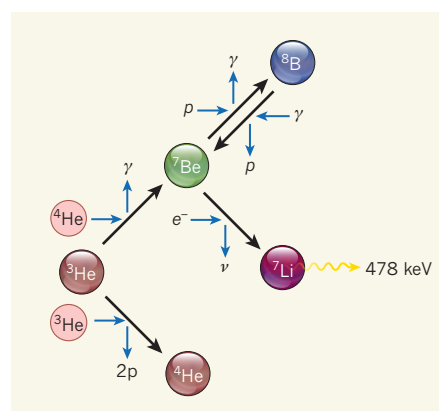


Figure 1 | The main nuclear reactions involved in the synthesis of ⁷Be and ⁷Li in novae. Tajitsu *et al.*¹ have detected radioactive ⁷Be in a nova explosion. ⁷Be transforms into ⁷Li, a neutrino and a photon of 478 kiloelectronvolts when it captures an electron. *p*, Proton; *e⁻*, electron; *γ*, *γ*-ray; ⁴He, helium-4; ³He, helium-3; ⁷Be, beryllium-7; ⁷Li, lithium-7; ⁸B, boron-8; *ν*, neutrino.

less abundant in the Milky Way and the Solar System than heavier elements.

Lithium has a complex origin. It is produced in three ways: by nucleosynthesis during the Big Bang; by nuclear reactions in the interstellar medium that are induced by energetic cosmic rays and are also responsible for the origin of beryllium and boron; and by nuclear reactions in stellar sources, such as red giants³. The stellar sources are required to reproduce the rise of lithium abundance in the Milky Way after the formation of the Solar System about 4.5 billion years ago.

Inside stars, ⁷Be, the subject of Tajitsu and colleagues' study, is formed by the fusion of helium-3 and helium-4. This radioactive element then captures an electron and transforms into its daughter nucleus, ⁷Li, within a short timescale (⁷Be has a half-life of 53.22 days), releasing a 478-kiloelectronvolt-energy photon (Fig. 1). But efficient production of ⁷Li requires this nuclear reaction to occur in hot, external stellar layers, and requires freshly produced ⁷Be to be transported into cooler subsurface layers before it transforms into ⁷Li. In this way, ⁷Li is immune to destruction once it is created. This process, known as the Cameron–Fowler ⁷Be transport mechanism, is responsible for ⁷Li production in stars^{4,5}.

Novae are thermonuclear explosions, and take place on top of white dwarfs that pull hydrogen-rich material from a companion star. As more hydrogen accumulates on the white dwarf, it builds up a shell that reaches pressures and temperatures sufficient to trigger explosive runaway fusion of the hydrogen. This leads to the fast expansion and subsequent ejection of the white dwarf's outer layers, and is accompanied by a sudden large increase in the star's brightness. During this process, ⁷Li is thought to be produced through the Cameron–Fowler ⁷Be transport mechanism.

The first studies of lithium production in novae were made in the 1970s^{6,7}, but it was not until 1996 that the details of the process were

pinned down⁸. It was realized that the initial chemical composition of the white dwarf that undergoes a nova was a crucial determinant of the amount of ⁷Li synthesized in the explosion; depending on the mass of its progenitor star, the white dwarf is made of either carbon and oxygen (CO novae) or oxygen and neon (ONe novae).

In CO novae, the carbon content makes the hydrogen fusion proceed faster than in ONe novae, owing to the operation of the CNO cycle of fusion reactions. Such faster evolution prevents the destruction of ³He and ⁷Be (Fig. 1), and so results in a larger production of ⁷Be and ⁷Li. The amount of ⁷Li produced by a CO nova corresponds to about 10⁻¹⁰ of the Sun's mass, but this value largely depends on the total ejected mass.

In their study, Tajitsu *et al.* report the detection of highly blue-shifted absorption lines of the singly ionized radioactive isotope of ⁷Be, ⁷Be II, in the near-ultraviolet spectra of the CO classical nova V339 Del, between 38 and 52 days after the explosion. The spectra were obtained using the Subaru Telescope of the

National Astronomical Observatory of Japan, which delivers high spectral resolution (about 0.0052 nanometres) and so allowed the authors to tease apart the lines of ⁷Be II from those of ⁹Be II, both of which occur at wavelengths around 312–313 nm.

The finding lends support to the hypothesis that the Cameron–Fowler ⁷Be transport mechanism is at work in novae, as predicted theoretically 40 years ago⁶. The observations indicate that nova V339 Del produced at least as much ⁷Be and ⁷Li as predicted by theory.

The implications of these results are manifold. First, they mean that novae may play a larger part in lithium production than previously thought. Second, they may increase the probability of detecting the 478-keV γ -ray photons emitted in the ⁷Be-to-⁷Li reaction⁹, which have remained elusive despite observational efforts made by γ -ray missions^{10,11}. Third, and perhaps most importantly, they suggest that measurements of ⁷Be lines in the near-ultraviolet range and within the lifetime of the element may well provide a way

of estimating the contribution of novae to the lithium abundance in the Milky Way and in the Universe in its entirety. ■

Margarita Hernanz is at the Institute of Space Sciences, ICE (CSIC-IEEC), 08193 Cerdanyola del Vallés, Barcelona, Spain.
e-mail: hernanz@ice.csic.es

1. Tajitsu, A., Sadakane, K., Naito, H., Arai, A. & Aoki, W. *Nature* **518**, 381–384 (2015).
2. Burbidge, E. M., Burbidge, G. R., Fowler, W. A. & Hoyle, F. *Rev. Mod. Phys.* **29**, 547–650 (1957).
3. Romano, D., Matteucci, F., Molaro, P. & Bonifacio, P. *Astron. Astrophys.* **352**, 117–128 (1999).
4. Cameron, A. G. W. *Astrophys. J.* **121**, 144–160 (1955).
5. Cameron, A. G. W. & Fowler, W. A. *Astrophys. J.* **164**, 111–114 (1971).
6. Arnould, M. & Nørgaard, H. *Astron. Astrophys.* **42**, 55–70 (1975).
7. Starrfield, S., Truran, J. W., Sparks, W. M. & Arnould, M. *Astrophys. J.* **222**, 600–603 (1978).
8. Hernanz, M., José, J., Coc, A. & Isern, J. *Astrophys. J.* **465**, L27–L30 (1996).
9. Clayton, D. D. *Astrophys. J.* **244**, L97–L98 (1981).
10. Harris, M. J., Leising, M. D. & Share, G. H. *Astrophys. J.* **375**, 216–220 (1991).
11. Harris, M. J. *et al. Astrophys. J.* **563**, 950–957 (2001).

EVOLUTION

Finches sequenced

Darwin's finches played a key part in the formulation of his theory of evolution by natural selection. They have since become an iconic model for adaptive radiation — 14 species evolved from a common ancestor to occupy different niches on the Galapagos Islands, with 1 species living on Cocos Island. On page 371 of this issue, Lamichhaney *et al.* present the genome sequences of 120 individuals from among all 15 species and 2 close relatives (S. Lamichhaney *et al. Nature* **518**, 371–375; 2015).

The work marks the first extensive genomic characterization of these birds. Unexpectedly, the analysis reveals that breeding between species has continued throughout their adaptation, contributing to their evolution.

The morphology of Darwin's finches has been extensively studied, with a particular focus on the diverse shapes of their beaks (pictured: *Geospiza magnirostris*). The authors use their rich data set to probe the genetic basis of beak shape, and identify six genomic regions that have a role in craniofacial morphology.

One region, which encodes the protein ALX1, is a major player in the rapid beak-shape evolution seen both across Darwin's finches and within one species, the medium ground finch (*Geospiza fortis*). The function of ALX1 is evolutionarily conserved — mutations in this gene also affect craniofacial development in humans and zebrafish. **Magdalena Skipper**



P. R. GRANT

BIOCHEMISTRY

Breaking methane

The most powerful oxidant found in nature is compound Q, an enzymatic intermediate that oxidizes methane. New spectroscopic data have resolved the long-running controversy about Q's chemical structure. SEE LETTER P.431

AMY C. ROSENZWEIG

Bacteria that consume methane gas (CH_4) to produce methanol (CH_3OH) using dioxygen (O_2) must break two chemical bonds: the bond holding the two oxygen atoms together, and one of the extremely strong carbon–hydrogen (C–H) bonds in methane. Knowing how these bonds are broken is central to the development of biological processes for converting methane into liquid fuels. Such processes offer a possible way of dealing with the methane that is wastefully burned or leaked to the atmosphere as a result of the worldwide hydraulic fracturing (fracking) boom. On page 431 of this issue, Banerjee *et al.*¹ report the chemical structure of the molecular species that reacts with methane in the active site of one of the enzymes that converts methane to methanol, soluble methane monooxygenase (sMMO).

Abundant and cheap natural gas is composed primarily of methane, and is a crucial source of fuel and chemicals. Unfortunately, large quantities of natural gas extracted together with oil are burned at some fracking sites to the tune of gas worth US\$100 million being wasted each month². Moreover, some of this methane is vented into the atmosphere, where it acts as a potent greenhouse gas. The problem could be alleviated by converting the wasted gas into liquid fuel at fracking sites, but gas-to-liquid (GTL) conversion of methane requires large-scale, expensive ‘Fischer–Tropsch’ facilities that are not easily established.

An alternative that has attracted much attention is biological GTL conversion using either bacteria that oxidize methane or isolated forms of the bacteria's primary metabolic enzyme, methane monooxygenase³ (MMO). Small-scale biological GTL facilities could be deployed at remote or temporary locations, and offer advantages over Fischer–Tropsch plants because GTL conversion occurs at ambient temperature and pressure; by contrast, Fischer–Tropsch chemistry requires high temperatures and pressures. But substantial increases in the rates of MMO reactions, as well as in the fractions of the carbon and energy present in methane that are converted to product (the carbon and energy efficiencies respectively) are necessary to create a viable technology³. Understanding the details of how MMOs work is germane to making such improvements.

There are two types of MMO: a membrane-bound, copper-containing enzyme⁴ (known as pMMO) and a soluble, iron-containing enzyme (sMMO). The latter belongs to a large family of bacterial multicomponent monooxygenases that use a pair of iron ions (a dinuclear iron centre) to oxidize hydrocarbons, but it is the only member that can oxidize methane⁵. Extensive studies⁶ over the past 20 years have worked out many details of the catalytic cycle of sMMO. First, the iron ions are reduced from the +3 oxidation state to the +2 state by a reductase protein. The dinuclear iron(II) centre then reacts with dioxygen in the presence of an essential regulatory protein to form peroxodiiron(III) intermediates. Next comes the key step: the oxygen–oxygen bond is cleaved, resulting in the formation of an intermediate called compound Q, which reacts with methane to break a C–H bond. Compound Q is then converted into a complex denoted T.

Compound Q has been investigated using a range of spectroscopic and computational approaches since it was first reported more than 20 years ago⁷. In 1997, Q was assigned a ‘diamond core’ structure consisting of two iron ions bridged symmetrically by two single

This experiment is challenging, but the authors overcame the obstacles using a specially designed instrument.

oxygen atoms⁸. This structure was proposed on the basis of data acquired using a technique called ⁵⁷Fe-Mössbauer spectroscopy, which indicated the presence of two iron(IV) ions occupying similar electronic and geometric environments, as well as X-ray absorption spectroscopic data that showed an unusually short iron–iron (Fe–Fe) distance (2.46 ångströms). However, computational work and studies of synthetic model compounds suggested longer Fe–Fe distances⁶ (2.6 to 2.8 Å), casting doubt on the proposed structure. By the late 2000s, the tide began to turn towards another possible structure, an ‘open core’ containing a terminal $\text{Fe}^{\text{IV}}=\text{O}$ unit — a motif found in model compounds that can oxidize C–H bonds rapidly⁹, although not those in methane.

In principle, the true nature of Q could be determined by resonance Raman spectroscopy, which can detect molecular vibrations from the stretching of iron–oxygen bonds; the frequencies of such vibrations provide a



50 Years Ago

In a written answer in the House of Commons on February 3, the Minister of Aviation, Mr. R. Jenkins, stated that a conference, attended by all the Member States of the European Launcher Development Organization, met in Paris during January 19–21, to review the activities of the organization and to examine proposals for its future work. The cost of completing the first programmed launcher... would be higher than the original estimate, and a working group had been set up to consider the feasibility of using the technical results already achieved and the experience acquired by the organization to develop an advanced launching system making use of the most modern techniques. The system would continue to use *Blue Streak* as the first stage, and the Conference would be convened again later this year to review progress.

From *Nature* 20 February 1965

100 Years Ago

In view of the shortage of fine chemicals which are used for research purposes, we thought it advisable to issue a circular to the principal laboratories in the kingdom asking for lists of chemicals not in immediate use, so that it would be possible to put the holders of such chemicals in touch with those chemists who were in urgent need of them. The replies which have been received so far have been in most cases to the effect that the holders wish to keep their own stocks in hand, but are willing to use our bureau for the purpose of purchasing others. As this attitude is one which entirely defeats the object with which the inquiry was started, may we direct the attention of chemists to the fact that it is impossible for them all to hold and to purchase at the same time.

From *Nature* 18 February 1915

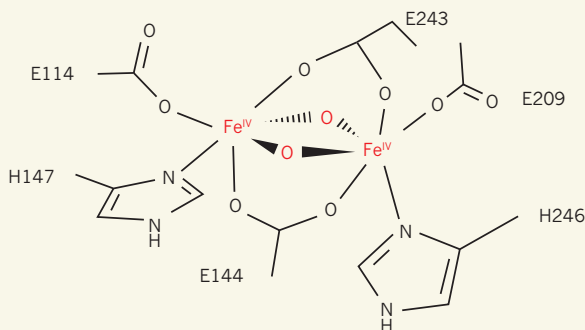


Figure 1 | The structure of Q. Banerjee *et al.*¹ report the structure of compound Q, a key intermediate in the conversion of methane to methanol by an enzyme called soluble methane monooxygenase. Their results indicate that Q's structure contains a 'diamond core' (red) in which two iron ions in the +4 oxidation state (Fe^{IV}) are bridged by oxygen atoms. The numbered groups in black surrounding the diamond core are the side chains of amino-acid residues. H, histidine residues; E, glutamate residues.

fingerprint for how the iron and oxygen atoms are bonded. This experiment is challenging for several reasons. First, intermediate Q forms only transiently, so the spectrum must be acquired in a time-resolved fashion. Second, signals from Q are expected to be weak because solutions of sMMO can be prepared at only low concentrations for analysis, and because of other experimental difficulties.

Banerjee *et al.* overcame these obstacles using a specially designed and optimized Raman instrument. In their set-up, a continuous stream of the diiron(II) enzyme was mixed with a second continuous stream of dioxygen-saturated buffer, and spectra were then acquired at different time points to capture the largest possible quantity of the short-lived Q. By comparing the spectra generated when both atoms in dioxygen were oxygen-18 isotopes ($^{18}\text{O}_2$) with those obtained using two oxygen-16 isotopes ($^{16}\text{O}_2$), they were able to isolate Q's vibration from the sea of other signals. A comparison of the frequency of this vibration with those observed for various iron-oxygen species in model complexes and enzymes gives only one match: the diamond core (Fig. 1). Importantly, the vibration does not correspond to a terminal $\text{Fe}^{\text{IV}}=\text{O}$ species, as would be expected for an open core structure.

To probe how the apparent diamond core forms, Banerjee and colleagues conducted experiments using a mixed isotopic form of dioxygen ($^{16}\text{O}-^{18}\text{O}$). They observed a new frequency in the spectrum of Q, which can be explained only by a diamond core that contains one ^{16}O and one ^{18}O atom, and which indicates that both atoms of dioxygen end up in Q. The spectra also reveal a vibration attributable to the product complex T, which contains one of the dioxygen atoms as a single unprotonated oxygen (an oxygen without a hydrogen atom attached) bridging the two iron ions. Further consideration of the results sheds light on how sMMO breaks the O–O bond to form intermediate Q. The data are most consistent with a mechanism in which the two electrons of the bond are distributed one to each oxygen

atom (homolytic cleavage), although it is not possible to completely rule out a mechanism in which both electrons go to the same oxygen atom (heterolytic cleavage).

Further verification of the Q structure is now desirable, and might be obtained from high-level computational studies and additional spectroscopic work. Diiron diamond cores have been previously observed in model complexes that cannot oxidize methane¹⁰, so what is it about Q that enables methane oxidation? One possibility suggested by Banerjee *et al.* is that a different

arrangement of the valence electrons of the iron(IV) ions in Q (a high spin state) confers increased reactivity, compared to the low spin state of synthetic complexes. This difference is probably just one of many ways that the enzyme micro-manages the oxidation chemistry to ensure Q's potency. ■

Amy C. Rosenzweig is in the Departments of Molecular Biosciences and of Chemistry, Northwestern University, Evanston, Illinois 60208, USA.
e-mail: amyr@northwestern.edu

1. Banerjee, R., Proshlyakov, Y., Lipscomb, J. D. & Proshlyakov, D. A. *Nature* **518**, 431–434 (2015).
2. Salmon, R. & Logan, A. *Flaring Up: North Dakota Natural Gas Flaring More than Doubles in Two Years* (Ceres, 2013); available at go.nature.com/jdks3y.
3. Haynes, C. A. & Gonzalez, R. *Nature Chem. Biol.* **10**, 331–339 (2014).
4. Culpepper, M. A. & Rosenzweig, A. C. *Crit. Rev. Biochem. Mol. Biol.* **47**, 483–492 (2012).
5. Sazinsky, M. H. & Lippard, S. J. *Acc. Chem. Res.* **39**, 558–566 (2006).
6. Tinberg, C. E. & Lippard, S. J. *Acc. Chem. Res.* **44**, 280–288 (2011).
7. Lee, S.-K., Nesheim, J. C. & Lipscomb, J. D. *J. Biol. Chem.* **268**, 21569–21577 (1993).
8. Shu, L. *et al. Science* **275**, 515–518 (1997).
9. Xue, G., De Hont, R., Munck, E. & Que, L. Jr *Nature Chem.* **2**, 400–405 (2010).
10. Xue, G. *et al. Proc. Natl Acad. Sci. USA* **104**, 20713–20718 (2007).

This article was published online on 21 January 2015.

CLIMATE SCIENCE

The future of coastal ocean upwelling

An ensemble of climate models predicts that winds along the world's coasts will intensify because of global warming, inducing more ocean upwelling — a process that will affect the health of coastal marine ecosystems. [SEE LETTER P.390](#)

EMANUELE DI LORENZO

At the ocean surface, where light is abundant, microscopic photosynthetic phytoplankton are the primary producers of organic material and the main source of energy for the oceanic food web. Phytoplankton growth depends on essential nutrients, which are typically depleted at the ocean surface but abundant in the deep ocean. Upwelling ocean currents carry these nutrients to the surface and thus support marine life. On page 390 of this issue, Wang *et al.*¹ report that many climate models predict that coastal upwelling will intensify in three of the most productive marine ecosystems of the world: the Canary, Benguela and Humboldt Eastern Boundary Upwelling Systems (EBUSs). This

result comes at a time when scientists are still debating the evidence supporting an increase in coastal upwelling, and its effects on coastal ecosystems and global carbon cycling.

Along the oceans' eastern boundaries, winds flowing along the coast drag surface waters out to sea. These displaced surface waters are replaced by water from lower down — the upwelling current. In 1990, the climate scientist Andrew Bakun realized that rising surface temperatures caused by the greenhouse effect would not be uniform: land will heat up faster than the ocean² (Fig. 1). Bakun proposed that this would create an ocean–land contrast in atmospheric pressure, which would drive stronger upwelling-favourable winds.

Wang and colleagues show that climate-model projections for the year 2100 support

Bakun's hypothesis by predicting a strong relationship between the strengthening of the land–ocean surface-temperature gradient and the intensification of the alongshore winds in most EBUSs. Furthermore, they find that this intensification will occur mostly at higher latitudes, where coastal upwelling is generally weaker. This in turn suggests that differences between the amount of upwelling at low and high latitudes will be reduced, causing homogenization of coastal upwelling habitats at different latitudes.

Is there observational evidence that winds have already increased along the coast? Scientists have debated this issue for the past 20 years, but there is a growing data-driven consensus that alongshore winds are indeed intensifying in EBUSs³. However, the future of coastal upwelling portrayed by Wang and co-workers comes with important caveats. Indices of coastal upwelling derived from alongshore winds are not the only indicators of upwelling strength⁴ and ecosystem impacts. The dynamics controlling the upward flux of nutrients (and therefore productivity) in the coastal ocean are complex and include processes that are not driven by alongshore winds.

For example, changes in upper-ocean stratification and deep-ocean nutrient concentrations⁵, changes in the energy of oceanic vortices⁶, extreme weather events and wind-stress gradients near the coast⁷ all affect coastal upwelling and marine ecosystems. Unfortunately, some of their effects are hard to predict in future scenarios of climate change, because they involve regional-scale ocean-transport dynamics that are not well represented in climate models.

Furthermore, upwelling systems undergo strong decadal climate-related fluctuations⁸, which might increase in amplitude as the climate changes, giving them a bigger effect than the long-term trends. Such decadal fluctuations occur in the California EBUS⁹, where Wang *et al.* found no significant increase in upwelling winds. Nevertheless, Wang and colleagues' study provides an invaluable starting point to think about the response of coastal upwelling systems to greenhouse forcing.

What are the potential ecological and societal impacts of increased coastal upwelling? It is estimated that phytoplankton growth in EBUSs already supports more than 20% of wild fisheries¹⁰. Most of this productivity occurs in the higher-latitude portions of the upwelling systems, where Wang and colleagues predict the strongest increase in upwelling. Increased upwelling in these regions might increase productivity and boost food production.

However, excessive productivity would generate heavier loads of organic matter that sinks into the deep ocean. Bacterial decomposition of this organic matter can deplete oxygen in the water column and, in extreme cases, generate deadly anoxic events at coastal upwelling sites¹¹. In the past few decades, coinciding with

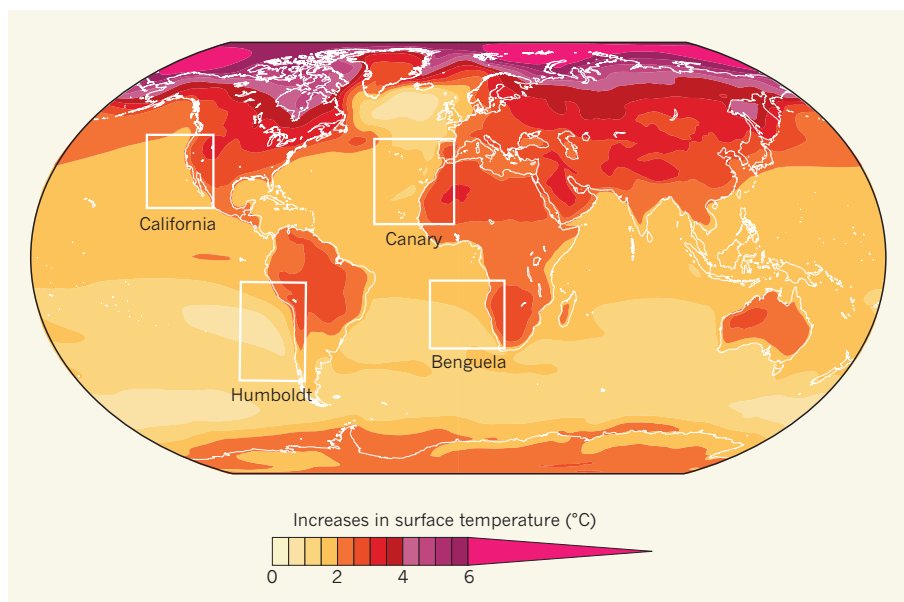


Figure 1 | Predicted changes in Earth's surface temperature. The map depicts differences in surface temperature between now and 2050, based on predictions from 27 climate models. The land heats up more than the ocean along major Eastern Boundary Upwelling Systems of the ocean (such as the California, Humboldt, Canary and Benguela systems). The resulting ocean–land temperature gradients have been hypothesized² to lead to stronger winds that favour upwelling along the coasts. Wang *et al.*¹ find support for this hypothesis using climate models. (Figure adapted from ref. 15.)

reports of oxygen depletion in ocean basins, these ecological 'dead zones' have become more apparent along coasts¹², raising concerns for the well-being of coastal ecosystems. Unfortunately, humans add to the risk by discharging heavy loads of nutrients along coasts, mostly as run-off of fertilizers from farmland.

Coastal upwelling also leads to degassing of carbon dioxide from deep water into the atmosphere. Surface ecosystems can offset

The authors' study provides an invaluable starting point to think about the response of coastal upwelling systems to greenhouse forcing.

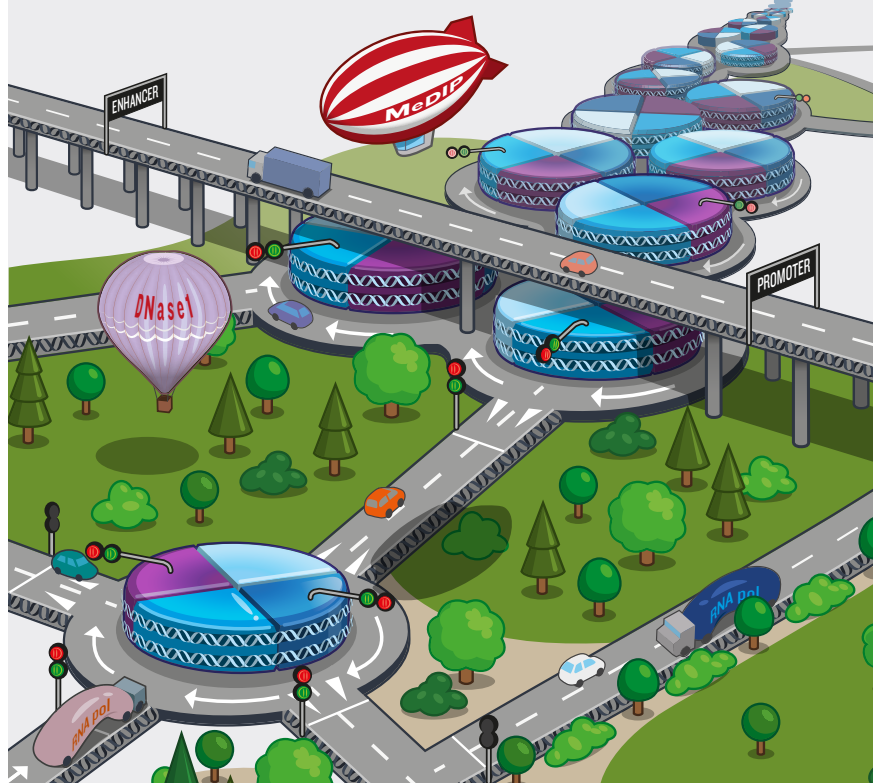
a rise of CO₂ degassing through increased photosynthesis, but the upwelling of carbon-rich water has other consequences. It is estimated that vertical mixing associated with oceanic physical processes has stored about half of the atmospheric CO₂ emitted since pre-industrial times in the deep ocean¹³. This has contributed to a progressive lowering of ocean pH and acidification of deep waters. Upwelling of these corrosive waters along the coasts has increasingly detectable effects on marine habitats and ecosystem functions¹⁴.

Increased upwelling currents will strongly affect marine ecosystems at EBUSs, but the long-term future of coastal acidification, dead zones and primary productivity probably depends on the properties of the water that comes to the surface. Observations and theories of deep ocean circulation show

that nutrient, oxygen and dissolved-carbon concentrations naturally undergo large fluctuations on timescales of decades to centuries. This variability is superimposed on climate trends, making it difficult to separate natural and anthropogenic contributions to changes in coastal marine ecosystems. Even so, it might be possible to use the slowly varying timescales of the deep ocean to make decadal predictions of acidification and hypoxia in upwelling areas. ■

Emanuele Di Lorenzo is at the School of Earth and Atmospheric Sciences, Georgia Institute of Technology, Atlanta, Georgia 30332-0340, USA. e-mail: edl@gatech.edu

1. Wang, D., Gouhier, T. C., Menge, B. A. & Ganguly, A. R. *Nature* **518**, 390–394 (2015).
2. Bakun, A. *Science* **247**, 198–201 (1990).
3. Sydeham, W. J. *et al. Science* **345**, 77–80 (2014).
4. Jacox, M. G., Moore, A. M., Edwards, C. A. & Fiechter, J. *Geophys. Res. Lett.* **41**, 3189–3196 (2014).
5. Rykaczewski, R. R. & Dunne, J. P. *Geophys. Res. Lett.* **37**, L21606 (2010).
6. Gruber, N. *et al. Nature Geosci.* **4**, 787–792 (2011).
7. Song, H., Miller, A. J., Cornuelle, B. D. & Di Lorenzo, E. *Dyn. Atmos. Oceans* **52**, 170–191 (2011).
8. Di Lorenzo, E. *et al. Oceanography* **26**, 68–81 (2013).
9. Sydeham, W. J., Santora, J. A., Thompson, S. A., Marinovic, B. & Di Lorenzo, E. *Glob. Change Biol.* **19**, 1662–1675 (2013).
10. Pauly, D. & Christensen, V. *Nature* **374**, 255–257 (1994).
11. Chan, F. *et al. Science* **319**, 920 (2008).
12. Grantham, B. A. *et al. Nature* **429**, 749–754 (2004).
13. Sabine, C. L. *et al. Science* **305**, 367–371 (2004).
14. Feely, R. A., Sabine, C. L., Hernandez-Ayon, J. M., Ianson, D. & Hales, B. *Science* **320**, 1490–1492 (2008).
15. Diefenbaugh, N. S. & Field, C. B. *Science* **341**, 486–492 (2013).



EPIGENOME ROADMAP

Epigenetics is implicated not just in the normal functioning of a cell and development but also in disease. Availability of the human genome sequence was a prerequisite for studies of genetic variation and its association with disease. The NIH Roadmap Epigenomics Program was created to provide a similar reference for genome-wide analyses of epigenetic changes — the epigenome.

The result is a public data resource containing information on DNA methylation, histone modifications, chromatin accessibility and small RNAs in stem cells and primary cell lines selected to represent tissues and organs frequently involved in disease.

This effort now culminates in the publication — here and elsewhere — of a considerable body of research papers.

On page 317 of this issue, the Roadmap Epigenomics Consortium describes the integrative analysis of 111 reference human epigenomes generated to establish global maps of regulatory elements. On page 350, Ren and colleagues describe the first human haplotype-resolved epigenomes, and show how they vary across tissues and among individuals.

Bernstein, Hafler and colleagues (page 337) present a fine-mapping strategy for genetic and epigenetic causal variants in 21 autoimmune diseases. On page 365, Tsai, Kellis and colleagues describe and compare hippocampal epigenomes of Alzheimer's disease patients with those of relevant mouse models. Stamatoyannopoulos, Sunyaev and colleagues (page 360) show how cell-of-origin chromatin organization shapes the mutational landscape of cancer.

Finally, three papers (pages 344, 355 and 331) dissect the role of epigenetic regulation in stem cell differentiation. Meissner and colleagues describe context-dependent rewiring of transcriptional regulation during differentiation of stem cells; Elkabetz, Meissner and colleagues studied regulatory networks during neural differentiation; while Ren and colleagues describe chromatin architecture changes during stem cell differentiation.

Research in this issue is accompanied by an online collection — the Epigenome Roadmap — which unites research from across Nature Publishing Group journals, as well as news stories and multimedia. Exploration of research papers is enhanced by 'threads', which highlight topics discussed in more than one paper.

We acknowledge the exclusive financial support of Illumina in producing this online collection. As always, *Nature* has full responsibility for all editorial content.

Magdalena Skipper Senior Editor (coordinating editor)
Alex Eccleston Senior Editor
Noah Gray Senior Editor
Therese Heemels Senior Editor

Nathalie Le Bot Senior Editor
Barbara Marte Senior Editor
Ursula Weiss Senior Editor

CONTENTS

NEWS & VIEWS

- 314 Roadmap for regulation**
C E Romanoski & C K Glass;
H G Stunnenberg; L Wilson & G Almouzni

ARTICLES

- 317 Integrative analysis of 111 reference human epigenomes**
Roadmap Epigenomics Consortium et al.
- 331 Chromatin architecture reorganization during stem cell differentiation**
J R Dixon et al.
- 337 Genetic and epigenetic fine mapping of causal autoimmune disease variants**
K K-H Farh et al.
- 344 Transcription factor binding dynamics during human ES cell differentiation**
A M Tsankov et al.

LETTERS

- 350 Integrative analysis of haplotype-resolved epigenomes across human tissues**
D Leung et al.
- 355 Dissecting neural differentiation regulatory networks through epigenetic footprinting**
M J Ziller et al.
- 360 Cell-of-origin chromatin organization shapes the mutational landscape of cancer**
P Polak et al.
- 365 Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease**
E Gjoneska et al.

MORE ONLINE

NATURE EPIGENOME ROADMAP

Epigenome Roadmap offers you a way of exploring a wealth of data from across Nature Publishing Group journals. By linking relevant paragraphs, figures and tables from all 24 papers in the collection, the 'threads' allow you to examine different themes.

www.nature.com/epigenomeroadmap



FORUM Epigenomics

Roadmap for regulation

A package of papers investigates the functional regulatory elements in genomes that have been obtained from human tissue samples and cell lines. The implications of the project are presented here from three viewpoints. [SEE ARTICLES P.317, P.331, P.337 & P.344 AND LETTERS P.350, P.355, P.360 & P.365](#)

THE TOPIC IN BRIEF

- Epigenomics is the study of the key functional elements that regulate gene expression in a cell.
- Epigenomes provide information about the patterns in which structures such as methyl groups tag DNA and histones (the proteins around which DNA is packaged to form chromatin), and about interactions between distant sections of chromatin.
- They also contain information about regulatory elements in DNA itself: both those that lie in the promoter region immediately upstream of where a gene's transcription begins, and those in distant enhancer sequences.

- The ENCODE Project¹ aimed to catalogue the regulatory elements in human cells, studying the epigenomic signatures of cells grown in culture. The Roadmap Epigenomics Project²⁻⁹ builds on this by analysing samples taken directly from human tissues and cells — embryonic and adult, diseased and healthy (Fig. 1).
- The researchers have linked these epigenomic data to the corresponding genetic information, producing reference epigenomes for 127 tissue and cell types.
- The result is a representation of how epigenomic elements regulate gene expression in the human body.

Four of the papers in this issue²⁻⁵ exploit these relationships to identify combinations of transcription factors that might define different cell types during development. Ziller *et al.*⁴ (page 355) modelled neuronal development *in vitro*, by generating six lineages of neuronal progenitors from embryonic stem (ES) cells, which give rise to almost every cell type of the body. The authors developed computational models to predict the transcription factors that bind to core neural-differentiation enhancers, as well as those that bind enhancers of distinct neural lineages only.

Tsankov *et al.*⁵ (page 344) studied the sets of transcription factors that bind to promoters and enhancers in the first three cell lineages that differentiate from ES cells. Sequences bound by transcription factors in one of the three lineages exhibited molecular modifications that promote gene expression, such as loss of DNA methylation. By contrast, the same DNA regions exhibited repressive modifications in the other two cell types. Both Ziller *et al.* and Tsankov *et al.* found that regulatory elements controlling genes that are essential for cellular identity are often also epigenetically modified in parental cells, highlighting the importance of existing regulatory landscapes and stage-specific expression of transcription factors for defining the developmental potential of cells.

Some major caveats should be noted. These studies are based on analysis of cell populations, and therefore miss potentially crucial aspects of cellular variability within populations. When tissues are examined, enhancer landscapes represent the composite of the cell types that make up that tissue, not a pure cell population. Studies^{10,11} of different populations of white blood cells called macrophages suggest that the tissue environment can shape enhancer landscapes, emphasizing the value of studying purified cell populations from *in vivo* sources. Finally, although the DNA sequences found in cell-specific enhancers provides clues

Differentiation enhanced

CASEY E. ROMANOSKI
& CHRISTOPHER K. GLASS

All the cells in the body contain essentially the same genome, and arise from the progeny of a single fertilized egg. How does each cell type interpret this common set of instructions to achieve its specific identity? The Roadmap Epigenomics Project has tackled this question by defining the epigenomic signatures of a broad spectrum of human tissues and cells undergoing crucial developmental transitions (for an overview², see page 317). Collectively, these papers and the associated data sets provide an unprecedented resource for understanding relationships between cells and tissues, and for delineating how cell-specific programs of gene expression are achieved.

Only about half of the approximately 25,000 protein-coding genes that make up the mammalian genome are expressed in any given cell type. Although many of these genes are required for general functions and are ubiquitously expressed, others are active in only one or a few cell types, or exhibit different patterns

of regulation from cell to cell. A remarkable achievement of the ENCODE Project was the use of epigenomic signatures to infer the existence of hundreds of thousands of enhancer-like regions in the mammalian genome that regulate gene expression at long range. From this vast palate, each cell type is regulated by a subset of perhaps 20,000–40,000 enhancers, which determine its particular gene-expression profile.

Enhancers are activated through interactions with transcription factors, which recognize and bind to specific DNA sequences within the enhancer region. Bound transcription factors recruit co-regulators, many of which deposit or remove modifications on histones. The way in which each cell type interprets genomic information is therefore closely linked to the organization of its DNA regulatory elements. Enhancers that are active in cell-type-specific epigenomic signatures are typically highly enriched in DNA sequences to which lineage-determining and signal-dependent transcription factors bind. Therefore, the delineation of a particular cell's active enhancer repertoire provides a powerful means of predicting the transcription factors required for that cell's identity. By extension, changes in epigenomic signatures during developmental transitions reflect activation or inhibition of such factors.



EPIGENOME ROADMAP

A Nature special issue
nature.com/epigenomeroadmap

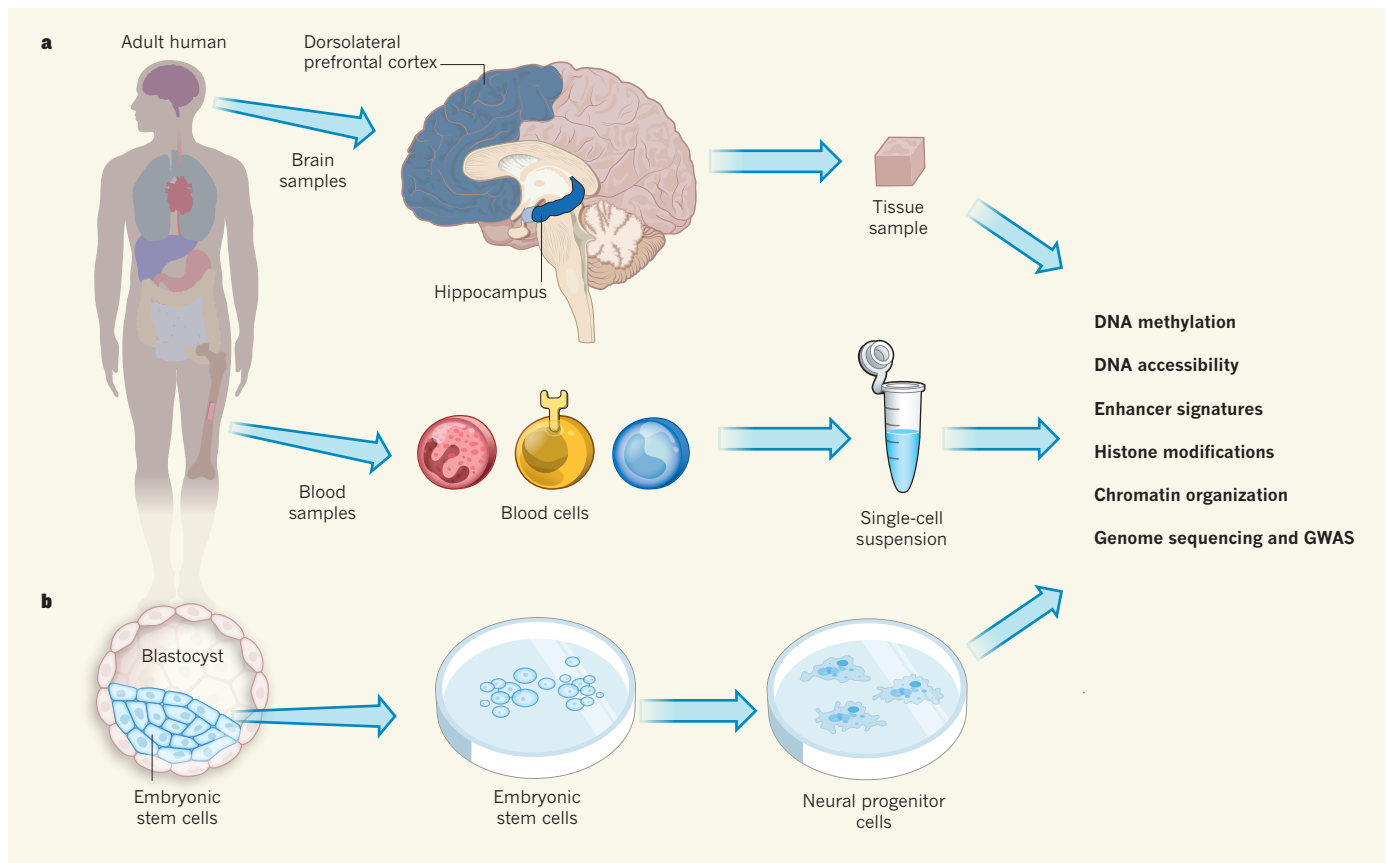


Figure 1 | From body to bench. The Roadmap Epigenomics Project has produced reference epigenomes that provide information on key functional elements controlling gene expression in 127 human tissues and cell types^{2–9}, and encompassing embryonic and adult tissues, from healthy individuals and those with disease. **a**, Many of the adult tissues investigated were broken down by cell type or region — blood into several types of immune cell, for instance, and the brain into regions including the hippocampus and dorsolateral

prefrontal cortex. Tissue samples and cells were subjected to a range of epigenomic analyses, along with genome sequencing and genome-wide association studies (GWAS). **b**, Embryonic stem (ES) cells, which are taken from the embryo at the ‘blastocyst’ stage and can give rise to almost every cell type in the body, were used to analyse, for example, the differentiation of stem cells into different neuronal lineages. The ES-cell-derived cell lines underwent the same epigenomic analyses as the tissue samples.

to the identities of the transcription factors that regulate enhancer activation, functional roles must be validated experimentally. The Roadmap Epigenomics Project has made some efforts along these lines, but the large number of hypotheses generated by the current papers means that this step is largely left for future work.

Casey E. Romanoski and Christopher K. Glass are in the Department of Cellular and Molecular Medicine, University of California, San Diego, La Jolla, California 92093-0651, USA.
e-mails: ckg@ucsd.edu; cromanoski@ucsd.edu

Diseases mapped

HENDRIK G. STUNNENBERG

For decades, biomedical science has focused on ways of identifying the genes that contribute to a particular trait, or phenotype. Approaches such as genome-wide association studies¹² (GWAS) identify locations in the

human genome at which variations in DNA sequence are linked to specific phenotypes, but if the variant is located in a region of DNA that does not encode a protein, such studies rarely provide insights into the regulatory mechanisms underlying the association. In these cases, comprehensive epigenomic analyses can provide the missing link between genomic variation and cellular phenotype.

The various consortia, including the Roadmap Epigenomics Program, that are gathered under the umbrella of the International Human Epigenome Consortium (www.ihc-epigenomes.org) have taken up the challenge of deciphering hundreds of cell-type-specific epigenomes using human cells and tissues from healthy donors and people with disease. In this issue, the Roadmap Epigenomics Project presents a wealth of epigenomes, a resource that provides a plethora of new hypotheses to be tested in relation to human health and disease. Given that epigenomes are cell-type specific, it makes sense to analyse disease-associated variants identified by GWAS in the context of the epigenome of the disease cell type. Indeed, previous groundbreaking observations¹³ revealed that non-protein-coding

genetic variants that are associated with phenotypic changes are often located in tissue-specific regulatory regions. The current papers use innovative analytical approaches to deepen and extend this knowledge.

Gjoneska *et al.*⁶ (page 365) made use of a mouse model of neurodegeneration that mimics Alzheimer’s disease. They found that disease-related changes in gene expression in the hippocampus of the mouse brain correlate with those in post-mortem brain samples taken from people with Alzheimer’s disease, but not with those from people without the disease. Subsequent detailed analyses revealed an upregulation of genes and regulatory regions linked to immune responses seen in Alzheimer’s disease. Genetic variants associated with the condition seemed to be enriched within evolutionarily conserved regulatory elements that control immune pathways, but not in neuronal pathways, providing fresh entry points for treatment.

Farh *et al.*⁷ (page 337) developed an algorithm to identify non-protein-coding genetic variants that might underlie autoimmune disease. The authors found that these variants are often located in or near enhancers

or promoters. However, only a small fraction of the variants cause a change in a sequence at which transcription factors are known to bind. This suggests that there is more to an enhancer than a 'simple' collection of sites of transcription-factor binding embedded in the composition of its DNA sequence. For example, flanking sequences might have a topological role affecting chromatin packaging and, consequently, DNA accessibility.

Polak *et al.*⁸ (page 360) investigated the distribution of cancer-associated genetic mutations in a set of diverse cancers, and correlated them with cell-type-specific epigenomic features. They found that the mutation profile of each cancer could often be predicted from the epigenomic signature of the cell type from which that cancer was most likely to have originated. Remarkably, the epigenomic signatures of cancer-cell lines (which are often used to study disease) were poor predictors of this profile. The authors conclude that the density and distribution of cancer mutations are strongly linked to a cell-type-specific epigenomic signature.

What comes next? The Roadmap Epigenomics Project has reached a major milestone, but the epigenomes of 127 cell types are just the beginning of the road to a comprehensive epigenome encyclopaedia. The International Human Epigenome Consortium plans to determine the epigenomes of every cell type in the human body — estimated to be several hundred to a thousand. Furthermore, each cell type must be analysed in many individuals, to assess the effect of genetic variation on personal cell-type-specific epigenomes. Finally, studies monitoring the epigenomic changes that arise as a result of ageing and of changes in environmental factors such as nutrients and metabolites will also be interesting. The epigenomics project has taught us that analysis and comparison of the genome and epigenome of healthy and diseased cells is essential for detecting and understanding the drivers of multifactorial diseases and traits.

Hendrik G. Stunnenberg is in the Department of Molecular Biology, Faculty of Sciences, Radboud University, Nijmegen 6525 GA, the Netherlands. e-mail: h.stunnenberg@ncmls.ru.nl

Chromatin charted

LAURENCE WILSON & GENEVIEVE ALMOUZZI

Chromatin is the complex of DNA, RNA and proteins that packages DNA within the cell. At the core of chromatin is an eight-subunit protein complex composed of histones. Molecular modifications to either DNA or

histones can affect the structure and function of chromatin. For example, some modifications promote chromatin compaction, affecting how easily DNA can be accessed by transcription factors, whereas others act as signals that modulate gene expression. A case in point is modification of the amino-acid residue lysine 27 (K27) on histone H3 in chromatin. Addition of an acetyl group (a modification known as H3K27ac) correlates with transcription of the corresponding region of DNA, whereas trimethylation (H3K27me3) is linked to transcriptional repression.

Several papers published by the Roadmap Epigenomics Project investigate histone modifications, and provide insights into the relationship between histone signatures and gene expression throughout development and adult life. For instance, three studies investigate the histone modifications associated with disease^{6–8}. Focusing on normal development, Tsankov *et al.*⁴ and Ziller *et al.*⁵ have mapped histone modifications that occur during the differentiation of embryonic cells (specifically, H3K4me1, H3K4me3, H3K27ac and H3K37me modifications), alongside patterns of transcription-factor binding and DNA methylation. They describe chromatin remodelling events that alter the accessibility of DNA sequences to which combinations of key regulatory transcription factors bind. These events correlate with the changes in gene expression that occur as cells differentiate.

In addition to the linear viewpoint of chromatin alterations presented through histone modifications, long-range chromatin interactions can also modulate gene expression — for instance, by bringing distant enhancers into contact with promoters that regulate the same gene. Dixon *et al.*⁹ (page 331) investigated this phenomenon, charting changes in three-dimensional (3D) chromatin organization during stem-cell differentiation. Human cells contain two copies, or alleles, of each gene, which can vary in terms of DNA sequence, resulting in differences in transcriptional activity (allele-restricted transcription). The allelic complement of a cell is known as its haplotype. Strikingly, Dixon and colleagues report that different haplotypes display different histone modifications and 3D chromatin organization, correlating with its allele-restricted transcription.

Leung *et al.*³ (page 350) confirmed this observation, reporting haplotype-specific differences in histone modifications and chromatin architecture that correlate with allele-restricted transcription across many tissues. Notably, these differences also correlate with mutations that disrupt sites of either transcription-factor binding or long-range chromatin interactions. However, the functional relevance of these imbalances remains to be deciphered.

These eight studies showcase the use of the first large-scale reference epigenome database,

taking advantage of the statistical power afforded by large sample sizes to formulate hypotheses about the relationships between the epigenome and the genome in different biological processes. They strengthen the link between chromatin modifications and gene expression in development and disease, defining core regulatory circuits that act in different tissues and at different developmental stages. This provides the community with a powerful reference tool, allowing researchers to compare the epigenome in their tissue of choice with snapshots from the database.

It is, however, still early days. Future work should try to address the changing relationship between the epigenome and genome over the lifespan of the cell, in different phases of the cell cycle and across cellular generations. Other factors that modulate chromatin organization also remain to be investigated — the proteins responsible for chromatin remodelling, for example, and the chaperone proteins associated with histone variants that control assembly and disassembly of chromatin¹⁴.

Defining the mechanisms that underlie chromatin-based regulation of gene expression will require integration of the observations made by the Roadmap Epigenomics Project with other approaches that directly test for function. For instance, model organisms will remain essential for comparative epigenomics and for garnering evolutionary information. Cutting-edge techniques, such as high-resolution microscopy, will allow live imaging of chromatin architecture and a means of studying its dynamics in space and time.

Above all, approaches and technologies that draw from different disciplines must be integrated in future epigenomic projects. This multidisciplinary approach is being catalysed by collaborations such as the EpiGeneSys network (www.epigenesys.eu), which bridges epigenetics and systems biology. Combining such efforts will be essential for understanding the functional link between the epigenome and the genome. ■

Laurence Wilson and Genevieve Almouzni are at the Institut Curie, CNRS Unit UMR3664, 75231 Paris Cedex 05, France. e-mail: genevieve.almouzni@curie.fr

1. The ENCODE Project Consortium. *Nature* **489**, 57–74 (2012).
2. Roadmap Epigenomics Consortium. *Nature* **518**, 317–330 (2015).
3. Leung, D. *et al.* *Nature* **518**, 350–354 (2015).
4. Ziller, M. J. *et al.* *Nature* **518**, 355–359 (2015).
5. Tsankov, A. M. *et al.* *Nature* **518**, 344–349 (2015).
6. Gjonneska, E. *et al.* *Nature* **518**, 365–369 (2015).
7. Farh, K. K.-H. *et al.* *Nature* **518**, 337–343 (2015).
8. Polak, P. *et al.* *Nature* **518**, 360–364 (2015).
9. Dixon, J. R. *et al.* *Nature* **518**, 331–336 (2015).
10. Gosselin, D. *et al.* *Cell* **159**, 1327–1340 (2014).
11. Lavin, Y. *et al.* *Cell* **159**, 1312–1326 (2014).
12. Welter, D. *et al.* *Nucleic Acids Res.* **42**, D1001–D1006 (2014).
13. Maurano, M. T. *et al.* *Science* **337**, 1190–1195 (2012).
14. Gurard-Levin, Z. A., Quivy, J.-P. & Almouzni, G. *Annu. Rev. Biochem.* **83**, 487–517 (2014).

Integrative analysis of 111 reference human epigenomes

Roadmap Epigenomics Consortium†, Anshul Kundaje^{1,2,3*}, Wouter Meuleman^{1,2*}, Jason Ernst^{1,2,4*}, Misha Bilenky^{5*}, Angela Yen^{1,2}, Alireza Heravi-Moussavi⁵, Pouya Kheradpour^{1,2}, Zhizhuo Zhang^{1,2}, Jianrong Wang^{1,2}, Michael J. Ziller^{2,6}, Viren Amin⁷, John W. Whitaker⁸, Matthew D. Schultz⁹, Lucas D. Ward^{1,2}, Abhishek Sarkar^{1,2}, Gerald Quon^{1,2}, Richard S. Sandstrom¹⁰, Matthew L. Eaton^{1,2}, Yi-Chieh Wu^{1,2}, Andreas R. Pfenning^{1,2}, Xinchun Wang^{1,2,11}, Melina Claussnitzer^{1,2}, Yaping Liu^{1,2}, Cristian Coarfa⁷, R. Alan Harris⁷, Noam Shores², Charles B. Epstein², Elizabeta Gjoneska^{2,12}, Danny Leung^{8,13}, Wei Xie^{8,13}, R. David Hawkins^{8,13}, Ryan Lister⁹, Chibo Hong¹⁴, Philippe Gascard¹⁵, Andrew J. Mungall⁵, Richard Moore⁵, Eric Chuah⁵, Angela Tam⁵, Theresa K. Canfield¹⁰, R. Scott Hansen¹⁶, Rajinder Kaul¹⁶, Peter J. Sabo¹⁰, Mukul S. Bansal^{1,2,17}, Annaick Carles¹⁸, Jesse R. Dixon^{8,13}, Kai-How Farh², Soheil Feizi^{1,2}, Rosa Karlic¹⁹, Ah-Ram Kim^{1,2}, Ashwinikumar Kulkarni²⁰, Daofeng Li²¹, Rebecca Lowdon²¹, GiNell Elliott²¹, Tim R. Mercer²², Shane J. Neph¹⁰, Vitor Onuchic⁷, Paz Polak^{2,23}, Nisha Rajagopal^{8,13}, Pradipta Ray²⁰, Richard C. Sallari^{1,2}, Kyle T. Siebenthal¹⁰, Nicholas A. Sinnott-Armstrong^{1,2}, Michael Stevens^{21,42}, Robert E. Thurman¹⁰, Jie Wu^{24,25}, Bo Zhang²¹, Xin Zhou²¹, Arthur E. Beaudet²⁶, Laurie A. Boyer¹¹, Philip L. De Jager^{2,23,27}, Peggy J. Farnham²⁸, Susan J. Fisher²⁹, David Haussler³⁰, Steven J. M. Jones^{5,31,32}, Wei Li³³, Marco A. Marra^{5,32}, Michael T. McManus³⁴, Shamil Sunyaev^{2,23,27}, James A. Thomson^{35,41}, Thea D. Tlsty¹⁵, Li-Huei Tsai^{2,12}, Wei Wang⁸, Robert A. Waterland³⁶, Michael Q. Zhang^{20,37}, Lisa H. Chadwick³⁸, Bradley E. Bernstein^{2,39,40§}, Joseph F. Costello^{14§}, Joseph R. Ecker^{9§}, Martin Hirst^{5,18§}, Alexander Meissner^{2,6§}, Aleksandar Milosavljevic^{7§}, Bing Ren^{8,13§}, John A. Stamatoyannopoulos^{10§}, Ting Wang^{21§} & Manolis Kellis^{1,2§}

The reference human genome sequence set the stage for studies of genetic variation and its association with human disease, but epigenomic studies lack a similar reference. To address this need, the NIH Roadmap Epigenomics Consortium generated the largest collection so far of human epigenomes for primary cells and tissues. Here we describe the integrative analysis of 111 reference human epigenomes generated as part of the programme, profiled for histone modification patterns, DNA accessibility, DNA methylation and RNA expression. We establish global maps of regulatory elements, define regulatory modules of coordinated activity, and their likely activators and repressors. We show that disease- and trait-associated genetic variants are enriched in tissue-specific epigenomic marks, revealing biologically relevant cell types for diverse human traits, and providing a resource for interpreting the molecular basis of human disease. Our results demonstrate the central role of epigenomic information for understanding gene regulation, cellular differentiation and human disease.

While the primary sequence of the human genome is largely preserved in all human cell types, the epigenomic landscape of each cell can vary considerably, contributing to distinct gene expression programs and biological functions^{1–4}. Epigenomic information, such as covalent histone modifications, DNA accessibility and DNA methylation can be interrogated in each cell and tissue type using high-throughput molecular assays^{2,5–8}. The resulting maps have been instrumental for annotating *cis*-regulatory elements and other non-exonic genomic features with characteristic epigenomic signatures^{9,10}, and for dissecting gene regulatory programs in development and disease^{7,9,11–14}. Despite these technological advances, we still lack a systematic understanding of how the epigenomic landscape contributes to cellular circuitry, lineage specification, and the onset and progression of human disease.

To facilitate and spearhead these efforts, the NIH Roadmap Epigenomics Program was established with the goal of elucidating how epigenetic processes contribute to human biology and disease. One of the major components of this programme consists of the Reference Epigenome Mapping Centers (REMCs)¹⁵, which systematically characterized the epigenomic landscapes of representative primary human tissues



EPIGENOME ROADMAP

A Nature special issue
nature.com/epigenomeroadmap

and cells. We used a diversity of assays, including chromatin immunoprecipitation (ChIP)^{9,10,16,17}, DNA digestion by DNase I (DNase)^{7,18}, bisulfite treatment^{1,2,19,20}, methylated DNA immunoprecipitation (MeDIP)²¹, methylation-sensitive restriction enzyme digestion (MRE)²², and RNA profiling⁸, each followed by massively parallel short-read sequencing (-seq). The resulting data sets were assembled into publicly accessible websites and databases, which serve as a broadly useful resource for the scientific and biomedical community. Here we report the integrative analysis of 111 reference epigenomes (Fig. 1 and Extended Data Fig. 1a–d), which we analyse jointly with an additional 16 epigenomes previously reported by the Encyclopedia of DNA Elements (ENCODE) project^{9,23}.

We integrate information about histone marks, DNA methylation, DNA accessibility and RNA expression to infer high-resolution maps of regulatory elements annotated jointly across a total of 127 reference epigenomes spanning diverse cell and tissue types. We use these annotations to recognize epigenome differences that arise during lineage specification and cellular differentiation, to recognize modules of regulatory regions with coordinated activity across cell types, and to identify key regulators of these modules based on motif enrichments and regulator

A list of affiliations appears at the end of the paper.

†Lists of participants and their affiliations appear at the end of the paper.

*These authors contributed equally to this work.

§These authors jointly supervised this work.

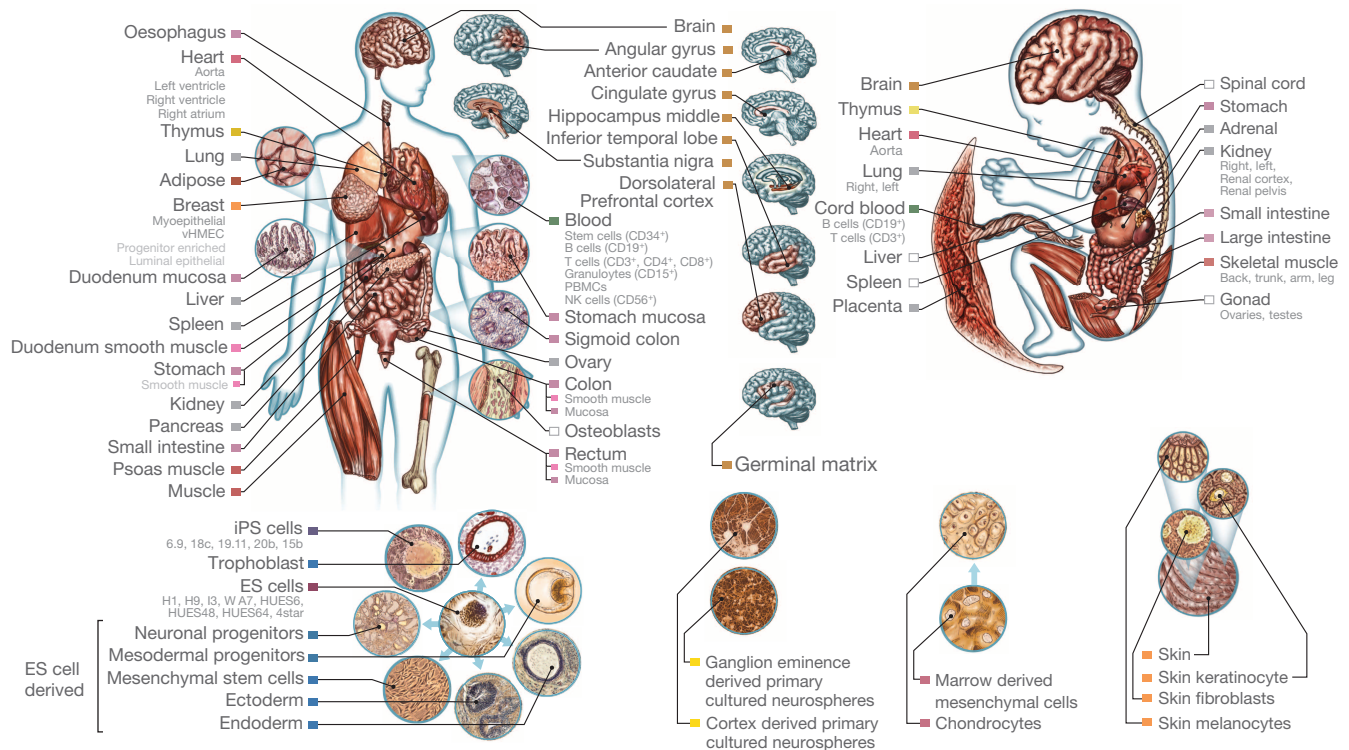


Figure 1 | Tissues and cell types profiled in the Roadmap Epigenomics Consortium. Primary tissues and cell types representative of all major lineages in the human body were profiled, including multiple brain, heart, muscle, gastrointestinal tract, adipose, skin and reproductive samples, as well as

expression. In addition, we study the role of regulatory regions in human disease by relating our epigenomic annotations to genetic variants associated with common traits and disorders. These analyses demonstrate the importance and wide applicability of our data resource, and lead to important insights into epigenomics, differentiation and disease. Specific highlights of our findings are given below.

- Histone mark combinations show distinct levels of DNA methylation and accessibility, and predict differences in RNA expression levels that are not reflected in either accessibility or methylation.
- Megabase-scale regions with distinct epigenomic signatures show strong differences in activity, gene density and nuclear lamina associations, suggesting distinct chromosomal domains.
- Approximately 5% of each reference epigenome shows enhancer and promoter signatures, which are twofold enriched for evolutionarily conserved non-exonic elements on average.
- Epigenomic data sets can be imputed at high resolution from existing data, completing missing marks in additional cell types, and providing a more robust signal even for observed data sets.
- Dynamics of epigenomic marks in their relevant chromatin states allow a data-driven approach to learn biologically meaningful relationships between cell types, tissues and lineages.
- Enhancers with coordinated activity patterns across tissues are enriched for common gene functions and human phenotypes, suggesting that they represent coordinately regulated modules.
- Regulatory motifs are enriched in tissue-specific enhancers, enhancer modules and DNA accessibility footprints, providing an important resource for gene-regulatory studies.
- Genetic variants associated with diverse traits show epigenomic enrichments in trait-relevant tissues, providing an important resource for understanding the molecular basis of human disease.

Reference epigenome mapping across tissues and cell types

The REMCs generated a total of 2,805 genome-wide data sets, including 1,821 histone modification data sets, 360 DNA accessibility data sets,

immune lineages, ES cells and iPS cells, and differentiated lineages derived from ES cells. Box colours match groups shown in Fig. 2b. Epigenome identifiers (EIDs, Fig. 2c) for each sample are shown in Extended Data Fig. 1.

277 DNA methylation data sets, and 166 RNA-seq data sets, encompassing a total of 150.21 billion mapped sequencing reads corresponding to 3,174-fold coverage of the human genome.

Here, we focus on a subset of 1,936 data sets (Fig. 2) comprising 111 reference epigenomes (Fig. 2a–d), which we define as having a core set of five histone modification marks (Fig. 2e). The five marks consist of: histone H3 lysine 4 trimethylation (H3K4me3), associated with promoter regions^{10,24}; H3 lysine 4 monomethylation (H3K4me1), associated with enhancer regions¹⁰; H3 lysine 36 trimethylation (H3K36me3), associated with transcribed regions; H3 lysine 27 trimethylation (H3K27me3), associated with Polycomb repression²⁵; and H3 lysine 9 trimethylation (H3K9me3), associated with heterochromatin regions²⁶. Selected epigenomes also contain a subset of additional epigenomic marks, including: acetylation marks H3K27ac and H3K9ac, associated with increased activation of enhancer and promoter regions^{27–29} (Fig. 2f); DNase hypersensitivity^{7,18}, denoting regions of accessible chromatin commonly associated with regulator binding (Fig. 2g); DNA methylation, typically associated with repressed regulatory regions or active gene transcripts^{4,30} and profiled using whole-genome bisulfite sequencing (WGBS)¹⁹, reduced-representation bisulfite sequencing (RRBS)²⁰, and mCRF-combined³¹ methylation-sensitive restriction enzyme (MRE)²² and immunoprecipitation based²¹ assays (Fig. 2h); and RNA expression levels⁸, measured using RNA-seq and gene expression microarrays (Fig. 2i). Our definition of 111 reference epigenomes is very similar to that used by the International Human Epigenome Consortium (IHEC), which required RNA-seq, WGBS and H3K27ac that are only available in a subset of epigenomes here. Lastly, an additional 16 histone modification marks on average were profiled across 7 deeply covered cell types (Fig. 2j).

We jointly processed and analysed our 111 reference epigenomes with 16 additional epigenomes from ENCODE^{9,23}. We generated genome-wide normalized coverage tracks, peaks and broad enriched domains for ChIP-seq and DNase-seq^{7,32}, normalized gene expression values for RNA-seq³³, and fractional methylation levels for each CpG site^{31,34,35}.

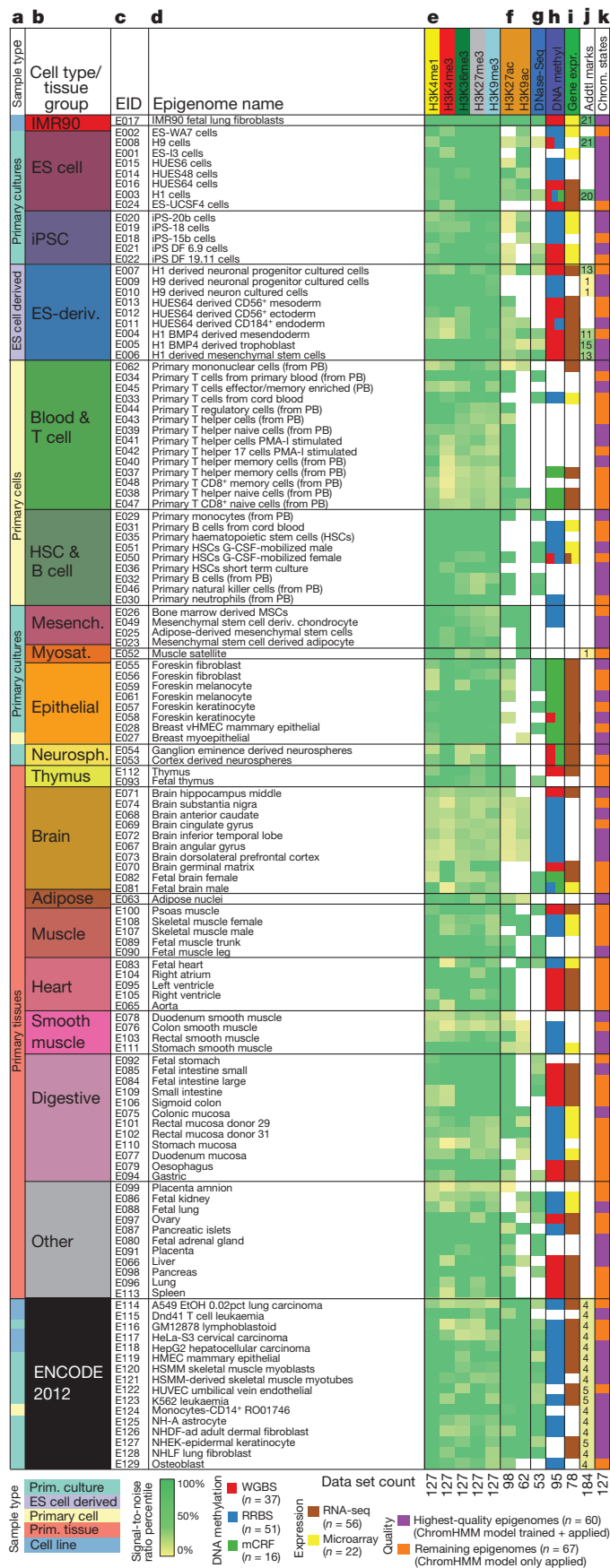


Figure 2 | Data sets available for each reference epigenome. List of 127 epigenomes including 111 by the Roadmap Epigenomics program (E001–E113) and 16 by ENCODE (E114–E129). See Supplementary Table 1 for a full list of names and quality scores. **a–d**, Tissue and cell types grouped by type of biological material (**a**), anatomical location (**b**), reference epigenome identifier (EID, **c**) and abbreviated name (**d**). PB, peripheral blood. ENCODE 2012 reference epigenomes are shown separately. **e–g**, Normalized strand cross-correlation quality scores (NSC)³⁷ for the core set of five histone marks (**e**), additional acetylation marks (**f**) and DNase-seq (**g**). **h**, Methylation data by WGBS (red), RRBS (blue) and mCRF (green). A total of 104 methylation data sets available in 95 distinct reference epigenomes. **i**, Gene expression data using RNA-seq (brown) and microarray expression (yellow). **j**, A total of 26 epigenomes contain 184 additional histone modification marks. **k**, Sixty highest-quality epigenomes (purple) were used for training the core chromatin state model, which was then applied to the full set of epigenomes (purple and orange).

genome-wide strand cross-correlation³⁷ (Fig. 2e–g); inter-replicate correlation; multidimensional scaling of data sets from different production centres (Supplementary Fig. 1); correlation across pairs of data sets (Extended Data Fig. 1e); consistency between assays carried out in multiple mapping centres (Supplementary Table 2); read mapping quality for bisulfite-treated reads^{38,39}; and agreement with imputed data⁴⁰. Outlier data sets were flagged, removed or replaced, and lower-coverage data sets were combined where possible (see Methods).

The resulting data sets provide global views of the epigenomic landscape in a wide range of human cell and tissue types (Fig. 3), including the largest and most diverse collection to date of chromatin state annotations (Fig. 3a); some of the deepest surveys of individual cell types using diverse epigenomic assays (with 21–31 distinct epigenomic marks for seven deeply profiled epigenomes; Fig. 3b); and some of the broadest surveys of individual epigenomic marks across multiple cell types (Fig. 3c). These data sets enable genome-wide epigenomic analyses across multiple dimensions (Fig. 3d). All data sets, standards and protocols are publicly available from web portals, linked from the main consortium homepage <http://www.roadmapepigenomics.org>, and also at <http://compbio.mit.edu/roadmap>.

Chromatin states, DNA methylation and DNA accessibility

As a foundation for integrative analysis, we used a common set of combinatorial chromatin states⁴¹ across all 111 epigenomes, plus 16 additional epigenomes generated by the ENCODE project (127 epigenomes in total), using the core set of five histone modification marks that were common to all. We trained a 15-state model (Fig. 4a, b and Supplementary Table 3a) consisting of 8 active states and 7 repressed states (Fig. 4c) that were recurrently recovered (Extended Data Fig. 2a), and showed distinct levels of DNA methylation (Fig. 4d), DNA accessibility (Fig. 4e), regulator binding (Extended Data Fig. 2b and Supplementary Fig. 2) and evolutionary conservation (Fig. 4f and Supplementary Fig. 3). The active states (associated with expressed genes) consist of active transcription start site (TSS) proximal promoter states (TssA, TssAFlnk), a transcribed state at the 5' and 3' end of genes showing both promoter and enhancer signatures (TxFlnk), actively transcribed states (Tx, TxWk), enhancer states (Enh, EnhG), and a state associated with zinc finger protein genes (ZNF/Rpts). The inactive states consist of constitutive heterochromatin (Het), bivalent regulatory states (TssBiv, BivFlnk, EnhBiv), repressed Polycomb states (ReprPC, ReprPCWk), and a quiescent state (Quies), which covered on average 68% of each reference epigenome. Enhancer and promoter states covered approximately 5% of each reference epigenome on average, and showed enrichment for evolutionarily conserved non-exonic regions⁴².

To capture the greater complexity afforded by additional marks, we trained additional chromatin state models in subsets of cell types. In the subset of 98 reference epigenomes that also included H3K27ac data, we also learned an 18-state model (Extended Data Fig. 2c and Supplementary Table 3b), enabling us to distinguish enhancer states containing strong H3K27ac signal (EnhA1, EnhA2), which showed higher DNA

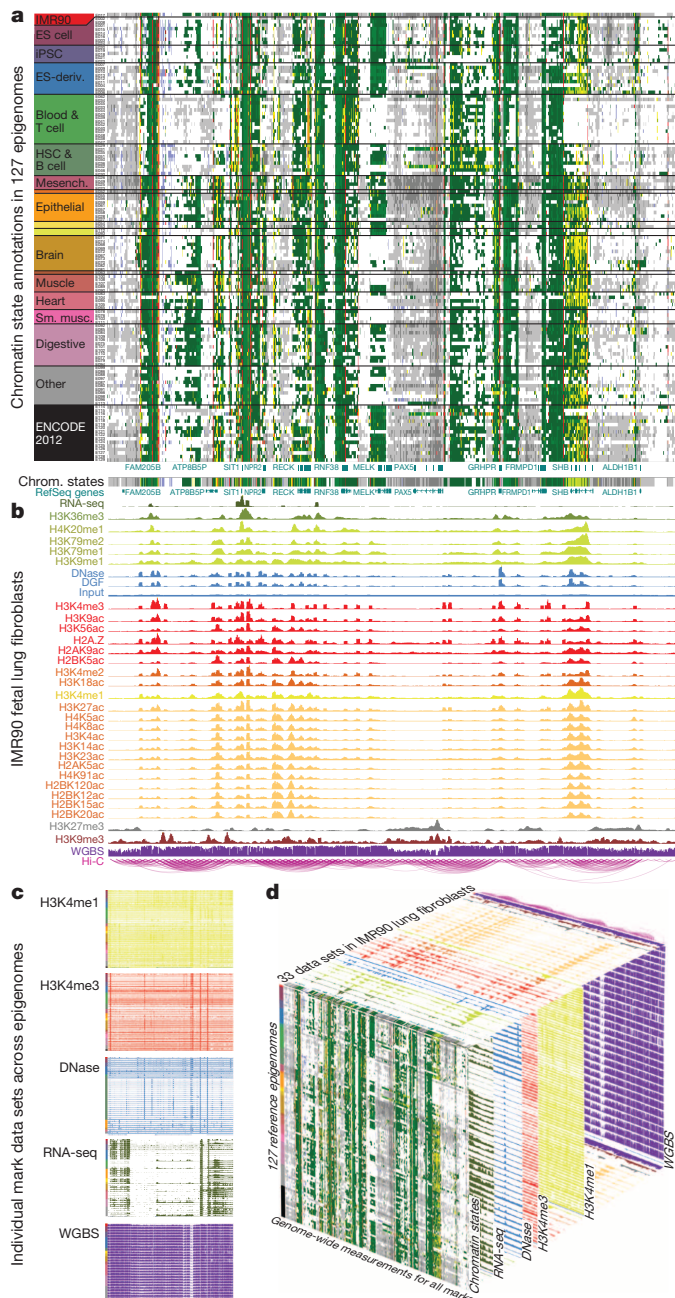


Figure 3 | Epigenomic information across tissues and marks. **a**, Chromatin state annotations across 127 reference epigenomes (rows, Fig. 2) in a ~3.5-Mb region on chromosome 9. Promoters are primarily constitutive (red vertical lines), while enhancers are highly dynamic (dispersed yellow regions). **b**, Signal tracks for IMR90 showing RNA-seq, a total of 28 histone modification marks, whole-genome bisulfite DNA methylation, DNA accessibility, digital genomic footprints (DGF), input DNA and chromatin conformation information⁷². **c**, Individual epigenomic marks across all epigenomes in which they are available. **d**, Relationship of figure panels highlights data set dimensions.

accessibility (Extended Data Fig. 3a), lower methylation (Extended Data Fig. 3b) and higher transcription factor binding (Extended Data Fig. 2c) than enhancers lacking H3K27ac. In a subset of 7 epigenomes with an average of 24 epigenomic marks, we learned separate 50-state chromatin state models based on all the available histone marks and DNA accessibility in each epigenome (Supplementary Fig. 4), which additionally distinguished: a DNase state with distinct transcription factor binding enrichments (Supplementary Fig. 4f), including for mediator/cohesin components⁴³ (even though CTCF was not included as an input

track to learn the model) and repressor NRSF; transcribed states showing H3K79me1 and H3K79me2 and associated with the 5' ends of genes and introns; and a large number of putative regulatory and neighbouring regions showing diverse acetylation marks even in the absence of the H3K4 methylation signatures characteristic of enhancer and promoter regions.

We used chromatin states to study the relationship between histone modification patterns, RNA expression levels, DNA methylation and DNA accessibility. Consistent with previous studies^{19,23,44,45}, we found low DNA methylation and high accessibility in promoter states, high DNA methylation and low accessibility in transcribed states, and intermediate DNA methylation and accessibility in enhancer states (Fig. 4d, e and Extended Data Fig. 3a, b). These differences in methylation level were stronger for higher-expression genes than for lower-expression genes, leading to a more pronounced DNA methylation profile (Extended Data Fig. 3c, Supplementary Fig. 5 and Supplementary Table 4f). Genes proximal to H3K27ac-marked enhancers show significantly higher expression levels (Extended Data Fig. 3d), and conversely, higher-expression genes were significantly more likely to be neighbouring H3K27ac-containing enhancers (Extended Data Fig. 3e).

Chromatin states sometimes captured differences in RNA expression that are missed by DNA methylation or accessibility. For example, TxFlnk, Enh, TssBiv and BivFlnk states show similar distributions of DNA accessibility but widely differing enrichments for expressed genes (Fig. 4c, d). Enh and ReprPC states show intermediate DNA methylation, but very different distributions of DNA accessibility and different enrichments for expressed genes (Fig. 4c–e). Lack of DNA methylation, typically associated with de-repression, is associated with both the active TssA promoter state and the bivalent TssBiv and BivFlnk states. Bivalent states TssBiv and BivFlnk also show overall lower DNA methylation and higher DNA accessibility than enhancer states Enh and EnhG, and binding by both activating and repressive regulatory factors (Extended Data Fig. 2b). These results also held for alternative methylation measurement platforms (Extended Data Fig. 4a–c), and for the 18-state chromatin state model (Extended Data Fig. 4d, e). Overall, these results highlight the complex relationship between DNA methylation, DNA accessibility and RNA transcription and the value of interpreting DNA methylation and DNA accessibility in the context of integrated chromatin states that better distinguish active and repressed regions.

Given the intermediate methylation levels of tissue-specific enhancer regions, we directly annotated intermediate methylation regions, based on 25 complementary DNA methylation assays of MeDIP^{31,46} and MRE-seq^{22,39} from 9 reference epigenomes⁴⁷. This resulted in more than 18,000 intermediate methylation regions, showing 57% CpG methylation on average, that are strongly enriched in genes, enhancer chromatin states (EnhBiv, EnhG, Enh) and evolutionarily conserved regions. Intermediate methylation was associated with intermediate levels of active histone modifications and DNase I hypersensitivity. Near TSSs, intermediate methylation correlated with intermediate gene expression, and in exons it was associated with an intermediate level of exon inclusion⁴⁷. Intermediate methylation signatures were equally strong within tissue samples, peripheral blood and purified cell types, suggesting that intermediate methylation is not simply reflecting differential methylation between cell types, but probably reflects a stable state of cell-to-cell variability within a population of cells of the same type.

Epigenomic differences during lineage specification

We next studied the relationship between DNA methylation dynamics and histone modifications across 95 epigenomes with methylation data, extending previous studies that focused on individual lineages^{19,48–50}. We found that the distribution of methylation levels for CpGs in some chromatin states varied significantly across tissue and cell type (Fig. 4g, Extended Data Fig. 4f and Supplementary Table 4a). For example, TssAFlnk states were largely unmethylated in terminally differentiated cells and tissues, but frequently methylated for several pluripotent and embryonic-stem-cell-derived cells (Bonferroni-corrected *F*-test *P* < 0.01);

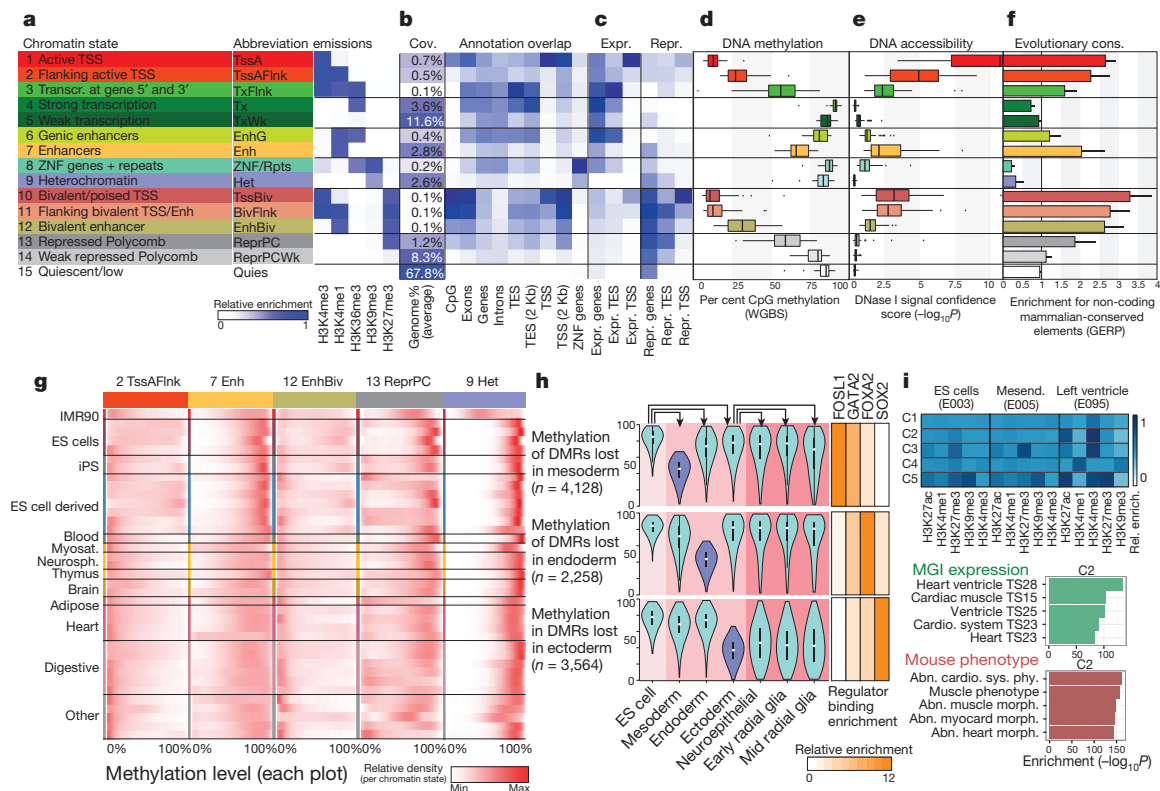


Figure 4 | Chromatin states and DNA methylation dynamics. **a**, Chromatin state definitions, abbreviations and histone mark probabilities. **b**, Average genome coverage. Genomic annotation enrichments in H1-ES cells. **c**, Active and inactive gene enrichments in H1-ES cells (see Extended Data Fig. 2b for GM12878). **d**, DNA methylation. **e**, DNA accessibility. **f**, Average overlap fold enrichment for GERP evolutionarily conserved non-exonic nucleotides. Bars

denote standard deviation. **g**, DNA methylation (WGBS) density (colour, ln scale) across cell types. Red = max ln(density + 1). Left column indicates tissue groupings; a full list is shown in Extended Data Fig. 4f. **h**, DNA methylation levels (left) and transcription factor enrichment (right) during ES cell differentiation^{50–53}. **i**, Chromatin mark changes during cardiac muscle differentiation. Heat map = average normalized mark signal in Enh. C2 cluster enrichment⁵⁵, with all clusters shown in <http://compbio.mit.edu/roadmap>.

Enh and EnhG states were highly methylated in pluripotent cells, but showed a broader distribution of intermediate methylation in differentiated cells and tissues ($P < 0.01$); EnhBiv states were unmethylated in most primary cells and tissues, but showed a broader distribution of methylation levels in pluripotent cells, possibly reflecting cell-to-cell heterogeneity ($P < 0.01$); the repressed state ReprPC showed varying methylation levels among epigenomes; and the Het state showed high levels of methylation in almost all epigenomes.

We also studied DNA methylation changes in three different systems. First, we studied DNA methylation changes during embryonic stem (ES) cell differentiation^{50,51}. We identified regions that lost methylation (differentially methylated regions (DMRs), Supplementary Table 4c) upon differentiation of ES cells (E003) to mesodermal (E013), endodermal (E011) and ectodermal (E012) lineages (Fig. 4h). Each lineage showed a largely distinct set of ~2,200–4,400 DMRs that are enriched for distinct transcription factor binding events (Fig. 4h, right column)⁵², consistent with their distinct developmental regulation. Upon further differentiation, ectodermal DMRs remained hypomethylated in three neural progenitor populations⁵³, despite the usage of distinct human ES cell (hESC) lines, and mesodermal and endodermal DMRs remained highly methylated (Fig. 4h), highlighting the lineage-specific nature of changes in DNA methylation during early differentiation^{50,54}.

Second, we studied DNA methylation changes associated with breast epithelia differentiation⁴⁵. Ectoderm to breast epithelia differentiation was dominated by DNA methylation loss (1.3M CpGs lost methylation compared with 0.2M gained), consistent with other primary somatic cell types⁵¹. By distinguishing luminal versus myoepithelial cells by flow sorting, and comparing a set of DMRs (Supplementary Table 4d) defined specifically in epithelial lineages⁴⁵, we found differences in nearest-gene enrichments⁵⁵ (mammary gland epithelium development versus

actin filament bundle, respectively) and differences in motif density (luminal DMRs show greater motif density for 51 transcription factors and lower density for 0 transcription factors). Proximal DMRs were highly associated with increased transcription, consistent with regulatory element de-repression associated with DNA methylation loss.

Third, we asked whether tissue environment or developmental origin is the primary driving factor in DNA methylation differences observed in more differentiated cell types⁵⁶ using epigenomes from skin cell types (keratinocytes E057/058, melanocytes E059/E061 and fibroblasts E055/056) that share a common tissue environment but possess distinct embryonic origins (surface ectoderm, neural crest and mesoderm, respectively). We found that despite the shared tissue environment, these three cell types displayed lower overlap in their DNA methylation and histone modification signatures, and instead were more similar to other cell types with a shared developmental origin. Using a set of DMRs (Supplementary Table 4e) defined specifically in the skin cell types⁵⁶, keratinocytes shared 1,392 (18%) of DMRs with surface ectoderm-derived breast cell types (hypergeometric P value $< 10^{-6}$), and 97% of these were hypomethylated. These shared DMRs were enriched for regulatory elements and cell-type-relevant genes, suggesting a common gene-regulatory network and shared signalling pathways and structural components⁵⁶. These results suggest that common developmental origin can be a primary determinant of global DNA methylation patterns, and sometimes supersedes the immediate tissue environment in which they are found.

We also examined coordinated changes in chromatin marks associated with cellular differentiation⁵⁷. We found that enhancers showing coordinated differences in multiple marks were enriched near genes showing common tissue-specific expression, and common knockout phenotypes based on their mouse orthologues. For example, enhancers

that showed higher H3K27ac and H3K4me3 (Fig. 4i, cluster C2) in left ventricle (E095) relative to their ES cells (E003) and mesodermal (E004) precursor lineages were enriched for heart ventricle expression and cardiac and muscle phenotypes in their mouse orthologues.

Most variable states and distinct chromosomal domains

We next sought to characterize the overall variability of each chromatin state across the full range of cell and tissue types. We first evaluated the observed consistency of each chromatin state at any given genomic position across all 127 epigenomes (Fig. 5a). We found that H3K4me1-associated states (including TxFlnk, EnhG, EnhBiv and Enh) are the most tissue specific, with 90% of instances present in at most 5–10 epigenomes, followed by bivalent promoters (TssBiv) and repressed states (ReprPC, Het). In contrast, active promoters (TssA) and transcribed states (Tx, TxWk) were highly constitutive, with 90% of regions marked in as many as 60–75 epigenomes. Quiescent regions were the most constitutive, with 90% consistently marked in most of the 127 epigenomes. These results held in the 18-state chromatin state model (Extended Data Fig. 5a), and in the subset of highest-quality epigenomes (Supplementary Fig. 6a, b).

Adjusting for the overall coverage and variability of each state, we then studied differences in the relative fraction of the genome annotated to each chromatin state between cell types (Fig. 5b, Extended Data Fig. 5b and Supplementary Fig. 6c–e). Haematopoietic stem cells and immune cells show a consistent and previously unrecognized depletion of active and bivalent promoters (TssA, TssBiv) and weakly transcribed states (TxWk), which may be related to their capacity to generate sub-lineages and enter quiescence (reversible G0 phase). ES cells and induced pluripotent stem cells (iPS cells) show enrichment of TssBiv, consistent with previous studies⁵⁸, and a depletion of ReprPCWk (defined by weak H3K27me3), possibly due to restriction of H3K27me3-establishing Polycomb proteins to promoter regions. Notably, IMR90 fetal lung fibroblasts, which were previously used as a somatic reference cell type⁵⁹, are

in fact a strong outlier in multiple ways, showing higher levels of Het, ReprPC and EnhG, and a depletion of Quies chromatin states.

We next studied the relative frequency with which different chromatin states switch to other states across different tissues and cell types (Fig. 5c), relative to switching in samples of the same tissue or cell type (Supplementary Fig. 7a, b). This revealed a relative switching enrichment between active states and repressed states, consistent with activation and repression of regulatory regions. The only exception was significant switching between transcribed states and active promoter and enhancer states, possibly due to alternative usage of promoters²² and enhancers⁶⁰ embedded within transcribed elements. These chromatin state switching properties were also found in the 18-state model incorporating H3K27ac marks (Extended Data Fig. 5c) and in the subset of 16 ENCODE reference epigenomes using both models (Supplementary Fig. 7c, d). We found that enhancers and promoters maintained their identity, except for a small subset of regions switching between enhancer signatures and promoter signatures⁶¹. Luciferase assays showed that these regions indeed possess both enhancer and promoter activity⁶¹, consistent with their epigenomic marks.

While chromatin states were defined at nucleosome resolution (200 bp), we also studied the overall co-occurrence of chromatin states across tissues at a larger resolution (2 Mb) to recognize higher-order properties (Fig. 5d). This analysis revealed that 2-Mb segments rich in active enhancers are constrained to approximately 40% of the genome (clusters c1–c6), with the remainder marked predominantly by inactive regions (c7–c11), consistent with the identification of two large chromatin conformation compartments^{12,62}. However, both compartments can be further subdivided by their chromatin state composition: inactive regions separate into predominantly quiescent (40%, c9, c11), heterochromatic (10%, c10), or bivalent (10%, c7, c8) marked regions; and active regions separate into regions rich in multiple marks (c3 and c6, showing a large diversity of active, ReprPC and bivalent states), enhancer and weakly transcribed regions (c5), and regions of intermediate activity (c1, c2, c4).

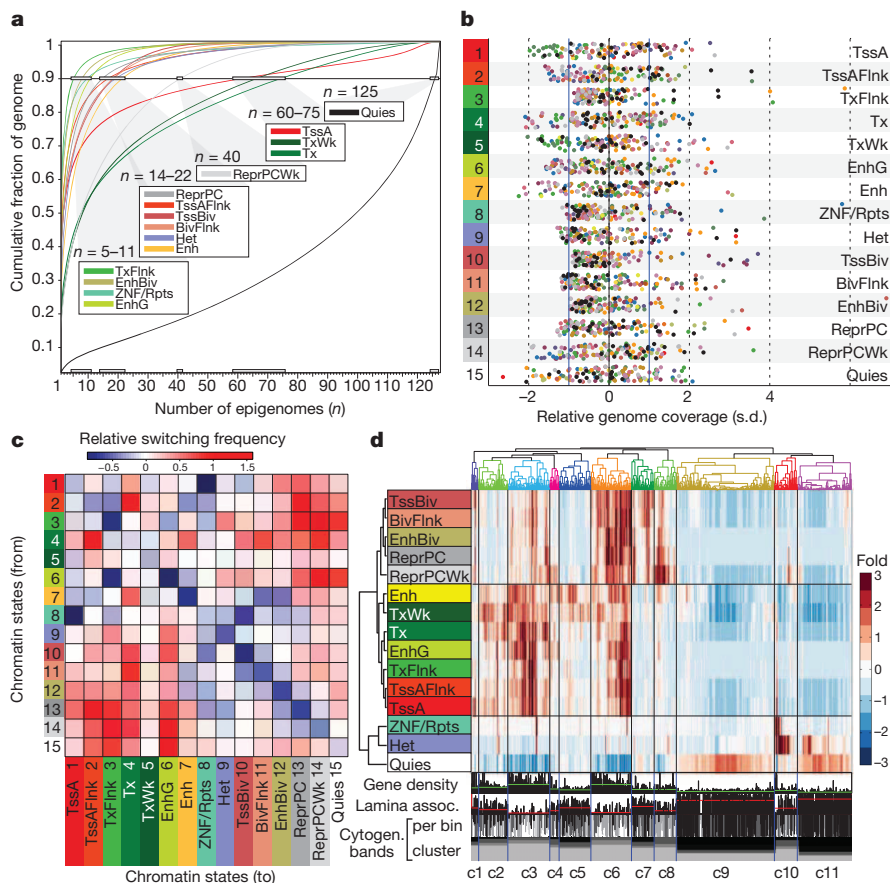


Figure 5 | Cell-type differences in chromatin states. **a**, Chromatin state variability, based on genome coverage fraction consistently labelled with each state. **b**, Relative chromatin state frequency for each reference epigenome. **c**, Chromatin state switching \log_{10} relative frequency (inter-cell type versus inter-replicate). **d**, Clustering of 2-Mb intervals (columns) based on relative chromatin state frequency (fold enrichment), averaged across reference epigenomes. LaminB1 occupancy profiled in ES cells. Red lines show cluster average.

These subdivisions were based on average state density across a large diversity of cell types and showed strong differences in gene density, CpG island occupancy, lamina association^{63,64} and cytogenetic bands (Fig. 5d and Extended Data Fig. 5d), suggesting that they represent stable chromosomal features.

Relationships between marks and lineages

We next studied the relationship between tissues and cell types, based on the similarity of diverse histone modification marks evaluated in their relevant chromatin states. Hierarchical clustering of our 111 reference epigenomes using H3K4me1 signal in Enh (Fig. 6a) showed consistent grouping of biologically similar cell and tissue types, including ES cells, iPSCs, T cells, B cells, adult brain, fetal brain, digestive, smooth muscle and heart. We also found several initially surprising but biologically meaningful groupings: fetal brain and germinal matrix samples clustered with neural stem cells rather than adult brain, consistent with fetal neural stem-cell proliferation; many ES-derived cells clustered with ES cells and iPSCs rather than the corresponding tissues, suggesting that those are still closer to pluripotent states than corresponding somatic states; adult and fetal thymus samples clustered with T cells rather than other tissues, consistent with roles in T-cell maturation and immunity. Several marks successfully recovered biologically meaningful groups when evaluated in their relevant chromatin states (Supplementary

Fig. 8), including H3K4me3 in TssA, H3K27me3 in ReprPC, and H3K36me3 in Tx, suggesting that the signal of each mark in relevant chromatin states is highly indicative of cell type and tissue identity. These alternative clusterings also showed some differences; for example, H3K4me3 in TssA states grouped several fetal samples together with each other, in a cluster neighbouring ES cells and iPSCs, rather than in separate tissue groups.

We applied this approach to compare the Roadmap Epigenomics reference epigenomes with the 16 ENCODE 2012 samples with broad mark coverage (Extended Data Fig. 6). We found that H3K4me1 signal in enhancer chromatin states correctly groups primary cells from similar tissues across the two projects, emphasizing the robustness of our annotations and signal tracks across projects (Extended Data Fig. 6a). For example, NHEK epidermal keratinocytes group with other keratinocytes, HMEC mammary epithelial cells group with other skin cells, and osteoblasts and HSMM skeletal muscle myoblasts group with bone marrow. Some cancer cell lines also grouped with corresponding primary tissues, including HepG2 hepatocellular carcinoma with liver tissue, NHLF primary lung fibroblasts with the IMR90 lung fibroblast cell line, and Dnd41 T-cell leukaemia with thymus, while in other cases cancerous cell lines grouped together, for example, HeLa-S3 cervical carcinoma with A549 lung carcinoma. H3K27me3 signal in Polycomb-repressed states grouped five immortalized cell lines together (Extended Data

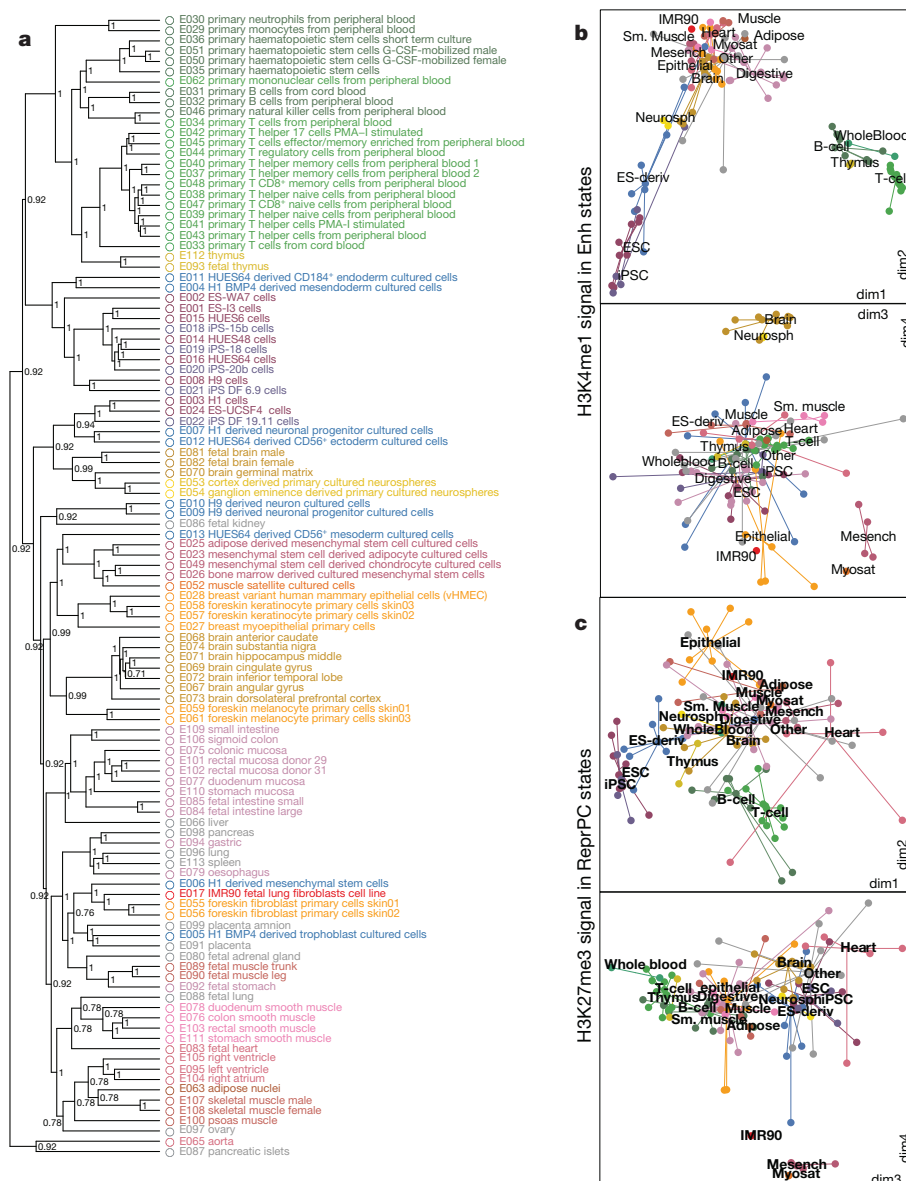


Figure 6 | Epigenome relationships.

a, Hierarchical epigenome clustering using H3K4me1 signal in Enh states. Numbers indicate bootstrap support scores over 1,000 samplings. **b, c**, Multidimensional scaling (MDS) plot of cell type relationships based on similarity in H3K4me1 signal in Enh states (**b**) and H3K27me3 signal in ReprPC states (**c**). First four dimensions are shown as dim1 versus dim2 and dim3 versus dim4.

Fig. 6c), despite their T-cell, lung, cervical, leukaemia and hepatocellular origins^{12,65}. The larger trees spanning ENCODE 2012 and Roadmap Epigenomics also highlighted the large number of lineages not previously covered by reference epigenomes, including brain, muscle, smooth muscle, heart, mucosa, digestive tract and fetal tissues.

To understand the relationship among different tissue/cell samples beyond the constraints of a tree representation, we also studied the full similarity matrix of each mark in relevant chromatin states (Supplementary Fig. 9) and also visualized the principal dimensions of epigenomic variation using multidimensional scaling (MDS) analysis (Supplementary Fig. 10). The pairwise similarity matrices of different marks were most effective in distinguishing different subsets of the samples, with H3K4me1 in Enh primarily capturing immune cell similarities, and H3K27me3 in ReprPC capturing pluripotent cell similarities (Supplementary Fig. 9). In the MDS analysis, the first four dimensions of variation for most marks separated major sample groups (Extended Data Fig. 7a–i), with some subtle differences between marks. For example, pluripotent cells and immune cells were two strong outliers in the first two dimensions of H3K4me1 variation in Enh (Fig. 6b), but H3K27me3 in ReprPC showed more uniform spreading of reference epigenomes (Fig. 6c), consistent with the coverage distributions of immune and pluripotent cells for the corresponding chromatin states (Fig. 5b). For most marks, the first five dimensions captured most of the variance, with additional dimensions capturing at most 4–6% for each mark (Extended Data Fig. 7).

Imputation and completion of epigenomic data sets

We exploited the strong relationships between marks and lineages for epigenomic signal imputation to complete missing marks across remaining tissues, and to complement observed data sets with more robust predictions based on multiple data sets⁴⁰. We predicted epigenomic signal tracks at 25-nucleotide resolution for histone marks, DNA accessibility, and RNA-seq data set and at single-base for CpG methylation, by exploiting correlations between multiple marks in the same cell type, and the same mark across multiple cell types.

We predict signal tracks for 34 epigenomic marks in 127 epigenomes, corresponding to 4,315 imputed genome-wide data sets, of which 3,193 (74%) are only available as imputed data. Imputed tracks showed high correlation with observed data, provided stronger and more consistent aggregate statistics relative to gene and TSS annotations, revealed lower-quality observed data sets in cases of disagreement between imputed and observed data, and captured cell type relationships and lineage-restricted information⁴⁰.

We also used 12 imputed epigenomic marks to learn a 25-state chromatin state model jointly across all 127 reference epigenomes, which distinguished multiple subtypes of enhancer and promoter regions across the complete set of reference epigenomes, including several active, weak and transcribed enhancer states, and both upstream and downstream promoter regions, providing an important reference annotation for studies of gene regulation and human disease⁴⁰.

Enhancer modules and their putative regulators

We next exploited the dynamics of epigenomic modifications at *cis*-regulatory elements to gain insights into gene regulation. We focused on 2.3M regions (12.6% of the genome) showing DNA accessibility in any reference epigenome and regulatory (promoter or enhancer) chromatin states, considering enhancer-only, promoter-only, or enhancer–promoter alternating states separately (Supplementary Fig. 11). We clustered enhancer-only elements (Enh, EnhBiv, EnhG) into 226 enhancer modules of coordinated activity (Fig. 7a), promoter-only elements into 82 promoter modules (Supplementary Fig. 11a) and promoter/enhancer ‘dyadic’ elements into 129 modules (Supplementary Fig. 11b), enabling us to distinguish ubiquitously active, lineage-restricted and tissue-specific modules for each group. Focusing on the enhancer-only clusters, we found that the neighbouring genes of enhancers in the same module showed significant enrichment for common functions⁶⁶ (Fig. 7b and

Supplementary Fig. 11c, d), common genotype–phenotype associations⁶⁷ (Fig. 7c), and common expression in their mouse orthologues (Supplementary Fig. 12), each annotation type showing strong consistency with the known biology of the corresponding tissues. For example, stem-cell enhancers are enriched near developmental patterning genes, immune cell enhancers near immune response genes, and brain enhancers near learning and memory genes (Fig. 7b). Sub-clustering of individual modules continued to reveal distinct enrichment patterns of individual sub-modules (Supplementary Fig. 11e), suggesting increased diversity of regulatory processes beyond the 226 modules used here.

The genome sequence of enhancers in the same module showed substantial enrichment for sequence motifs⁶⁸ associated with diverse transcription factors (Supplementary Fig. 13a). We found 84 significantly enriched motifs in 101 modules (Extended Data Fig. 8), indicating that enhancer modules likely represent co-regulated sets, and proposing candidate upstream regulators for nearly half of all modules. Direct application of the same approach and thresholds to the putative regulatory regions annotated in each of the 111 reference epigenomes led to significant enrichment for only 10 enriched motifs in 15 reference epigenomes (Supplementary Fig. 13b, c) of which 8 are blood samples, and focusing on the regions unique to each of the 17 tissue groups (Fig. 2b) only led to 19 enriched motifs in 10 tissue groups (Supplementary Fig. 13d, e), emphasizing the importance of studying regulatory motif enrichments at the level of enhancer modules.

We next sought to distinguish likely activator and repressor motifs, by identifying regulators with expression patterns across cell/tissue types that show a strong (positive or negative) correlation with the activity of enhancers in the corresponding modules⁹. We focused on the 40 most strongly expression-correlated regulators (Extended Data Fig. 9a), and used the module-level motif enrichments to link each regulator to the cell/tissue types that define each module (Fig. 8). We found that many of the inferred links correspond to known regulatory relationships, including OCT4 (also known as POU5F1) in pluripotent cells, HNF1B and HNF4A1 in liver and other digestive tissues, RFX4 in neurosphere and neuronal cells, and MEF2D in muscle. The most enriched regulators showed primarily positive correlations, suggesting that they function as transcriptional activators, while a subset of factors showed a negative correlation, with the motif showing enhancer depletion in the lineages where the corresponding factor is expressed, suggesting a repressive role. For example, REST (also known as NRSF), a known repressor of neuronal lineages, showed lowest expression in neuronal tissues, where its motif was most enriched in enhancers, and a similar signature was found for ZBTB1B, a known repressor of myogenesis and brain development.

Regulatory motifs predicted to be drivers of enhancer activity patterns showed significant enrichment in tissue-specific high-resolution (6–40 bp) DNase digital genomic footprints (DGF)⁶⁹ in matching cell types (Extended Data Fig. 9b and Supplementary Table 5b), providing DNA accessibility evidence that the motifs are indeed bound in these cell types. In addition, they showed positional bias relative to both the centre of DGF locations and relative to their boundaries (Extended Data Fig. 10), a property not found for shuffled motifs⁷⁰. These positional biases were highly tissue- and cell-type-specific for most activating factors (Extended Data Fig. 9c), including POU5F1 in iPS cells, MEF2D in heart, HNF1B in gastrointestinal tissues, BHLH in brain, SPI1 in immune cells, and MEF2 in heart and muscle, in each case matching the tissues that showed the highest enrichment. In contrast, for repressive factors and CTCF, positional biases were found in large numbers of tissues, even when the motifs were not enriched in active enhancers. For example, REST (NRSF) was positionally biased in DGF sites in nearly all tissues except brain (Extended Data Fig. 9c), even though it was only enriched in active enhancers in brain (Extended Data Fig. 9a), consistent with widespread repressive binding in non-brain tissues.

Overall, these enhancer modules, motif enrichments and regulatory predictions provide an unbiased map that can help guide studies

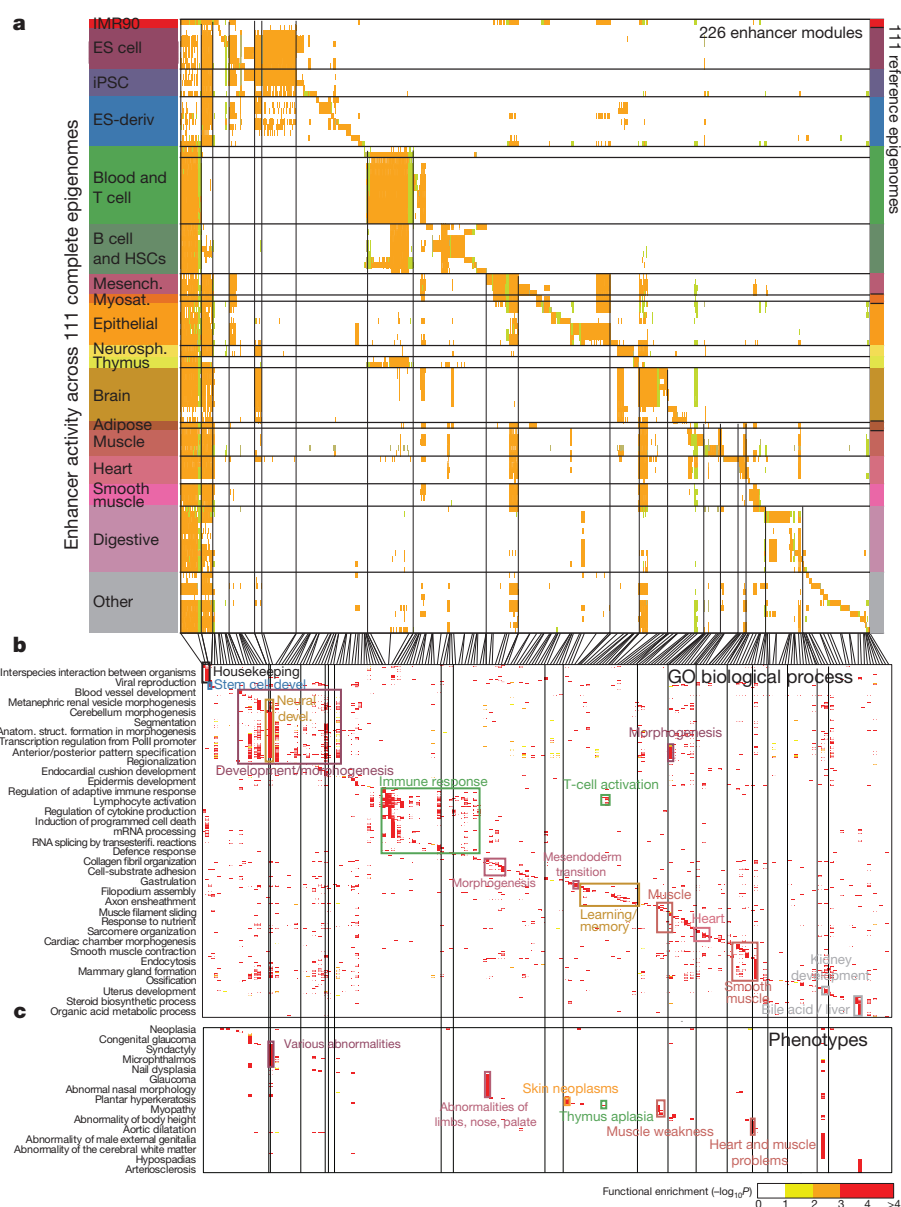


Figure 7 | Regulatory modules from epigenome dynamics. **a**, Enhancer modules by activity-based clustering of 2.3 million DNase-accessible regions classified as Enh, EnhG or EnhBiv (colour) across 111 reference epigenomes. Vertical lines separate 226 modules. Broadly active enhancers are shown first. Module IDs are shown in Supplementary Fig. 11c. **b, c**, Proximal

gene enrichments⁵⁵ for each module using gene ontology (GO) biological process (**b**) and human phenotypes (**c**). Rectangles pinpoint enrichments for selected modules. Representative gene set names (left) were selected using bag-of-words enrichment.

of candidate master regulators for fetal and adult lineage establishment and cell-type identity.

Impact of DNA sequence and genetic variation

We next studied the impact of primary DNA sequence on the epigenomic landscape, across genomic regions and between the two alleles of a given individual. First, we evaluated whether histone modifications and DNA methylation can be predicted by the underlying DNA sequence using DNA motifs for transcription factors expressed in ES cells and four ES-derived cell types. Using the area under the receiver operating curve (AUROC), we found between 71% predictive power for H3K4me1 peaks and 98% for H3K4me3 peaks (average of 85% across six marks and methylation-depleted regions)⁷¹. The most predictive motifs were those of factors associated with specific histone modifications or specific cell types, and were found within peak regions enriched for chromatin marks and at their boundaries. As an example of a boundary enrichment, H3K4me3 peaks were flanked by motifs consisting

of a continuous stretch of A and T followed by a G and C, which may have a role in nucleosome positioning or recruiting promoter-associated transcription factors, such as nuclear receptors. Enhancer and promoter-predictive motifs were enriched in high-resolution DNase hypersensitive sites (Supplementary Table 5a), suggesting that they correspond to transcription-factor-bound sequences.

Second, we studied how sequence variants between the two alleles of the same individual can lead to allelic biases in histone modifications, DNA methylation and transcript levels. We reconstructed chromosome-spanning haplotypes for ES cells, four ES-cell-derived cell lines⁷² and 20 tissue samples⁶¹, and we resolved allele-specific activity and structure for each. We found widespread allelic bias in both transcript levels and epigenomic marks for each epigenome. For example, 24% of all testable genes that contain exonic variants demonstrate allelic transcription in one or more ES cell or ES-cell-derived cell lineages, and the majority of these genes also exhibit allelic epigenomic modifications in promoters (71%) and Hi-C-linked enhancers (69%)⁷². Similarly, as many as 11%

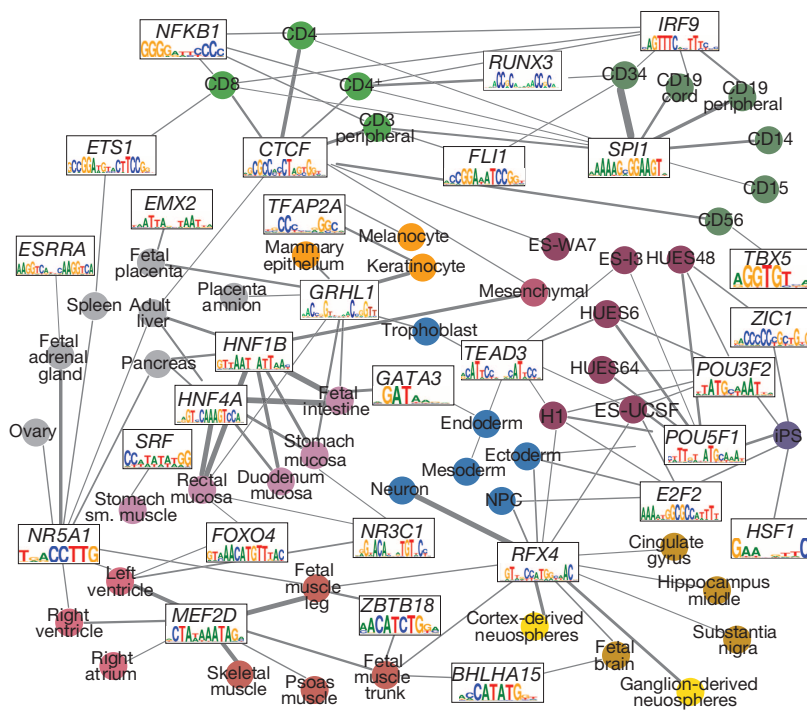


Figure 8 | Linking regulators to target tissues and cell types. Module-level regulatory motif enrichment (Supplementary Fig. 11) and correlation between regulator expression and module activity patterns (Extended Data Fig. 8a) are used to link regulators (boxes) to their likely target tissue and cell types (circles). Edge weight represents motif enrichment in the reference epigenomes of highest module activity.

of the testable enhancers display allelic bias in histone modification H3K27ac in the 20 tissue samples with allele-resolved transcription and chromatin states⁶¹. Allelic histone acetylation at enhancers is highly specific to individual genotypes, and often occurs near sequence variants that alter transcription factor binding, suggesting *cis*-acting sequence drivers for at least a subset of these regions^{61,72}.

Trait-associated variants enrich in tissue-specific marks

We next used our tissue-specific epigenomic data sets to study the regulatory annotation enrichments of phenotype-associated variants from genome-wide association studies (GWAS) of diverse traits and disorders. Previous studies showed that disease-associated variants are enriched in specific regulatory chromatin states⁹, evolutionarily conserved elements⁷³, histone marks⁷⁴ and accessible regions¹⁴. We expanded these analyses using the diversity of primary tissues surveyed by our epigenomic maps, applied to a compendium of disease-associated variants from the NHGRI GWAS catalogue⁷⁵. We intersected the set of variants identified in each curated study with peaks of H3K4me1, H3K4me3, H3K36me3, H3K27me3 and H3K9me3 across each of the 127 epigenomes, and H3K27ac, H3K9ac and DNase when available (Extended Data Figs 11, 12 and Supplementary Table 6), and we searched for significant enrichment in their overlap relative to what would be expected given the NHGRI GWAS catalogue as background (see Methods).

For enhancer-associated H3K4me1 peaks, we found 58 studies (Fig. 9a and Extended Data Fig. 11a) with significant enrichments in at least one tissue at 2% false discovery rate (FDR) (hypergeometric $P < 10^{-3.9}$). Upon manual curation, the enriched cell types were consistent with our current understanding of disease-relevant tissues for the vast majority of cases. For example, diverse immune traits were enriched in immune cell enhancers, including rheumatoid arthritis, coeliac disease, type 1 diabetes, systemic lupus erythematosus, chronic lymphocytic leukaemia, allergy, multiple sclerosis, and Graves' disease^{76–82}. A large number of metabolic trait variants are enriched in liver enhancer marks, including LDL, HDL, total cholesterol, lipid metabolism phenotypes, and metabolite levels^{83,84}. Fasting glucose was most enriched for pancreatic islet enhancer marks and insulin-like growth factors in placenta, consistent with their endocrine regulatory roles^{85,86}. Several cardiac traits were enriched in heart tissue enhancers, including the PR heart repolarization interval, blood pressure and aortic root size. Interestingly, inflammatory bowel disease and ulcerative colitis variants show

enrichment in both immune and gastrointestinal enhancer marks, suggesting that dysregulation of both organs may underlie disease predisposition. Both attention deficit hyperactivity disorder and adiponectin levels were enriched in brain regions, consistent with causal roles in brain dysregulation^{87,88}. In contrast, late-onset Alzheimer's disease variants were enriched in immune cell enhancers, rather than brain, consistent with recent evidence of a possible immune and inflammatory basis^{89–91}.

For active enhancer-associated H3K27ac peaks (available in 98 cell types), we found a similar number of enriched studies (47 at 2% FDR, Extended Data Fig. 12b), but for promoter-associated H3K4me3 and H3K9ac peaks, we found only 25 and 18 enriched studies, respectively (Extended Data Fig. 12a, b), suggesting that enhancer-associated marks are more informative for tissue-specific disease enrichments than promoter-associated marks. For DNase peaks, we only found 9 enriched studies (Extended Data Fig. 12c), partly because they were only available in 53 reference epigenomes (restricting H3K4me1 to the same 53 resulted in 25 enriched studies, Supplementary Table 6), and possibly due to lack of distinction between enhancer and promoter regions. For transcription-associated H3K36me3, we found 15 enriched studies (Extended Data Fig. 12d), indicating that these help capture additional biologically meaningful variants outside annotated promoter and enhancer regions. In contrast, we found no enriched study for either Polycomb-associated H3K27me3 peaks or heterochromatin-associated H3K9me3 peaks (Extended Data Fig. 12e, f). These results indicate that enhancer-associated marks have the greatest ability to distinguish tissue-specific enrichments for regulatory regions, but promoter-, open-chromatin- and transcription-associated marks also have numerous significant enrichments, suggesting that disease variants affect a wide range of processes.

These results illustrate that the epigenomic annotations provided here across a broad range of primary tissues and cells will be of great utility for interpreting genetic changes associated with complex traits. We have made all these epigenomic annotations of GWAS regions publicly searchable and browsable through the Roadmap Epigenome Browser⁹² and an updated version of the HaploReg database⁹³.

Discussion

The NIH Roadmap Epigenomics Program has been working to improve epigenomic assays, generate reference epigenomic maps, and use them to understand gene regulation, differentiation, reprogramming and

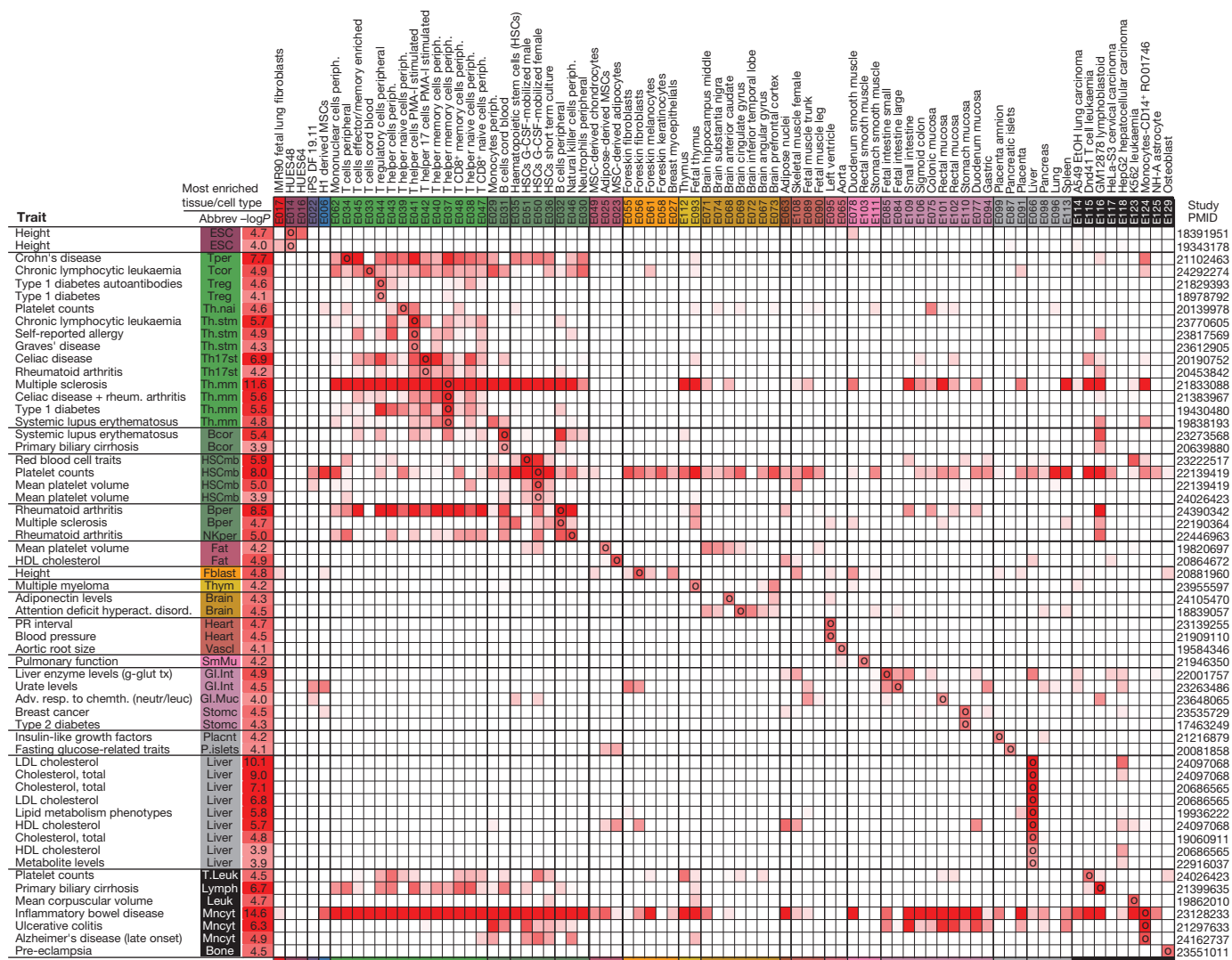


Figure 9 | Epigenomic enrichments of genetic variants associated with diverse traits. Tissue-specific H3K4me1 peak enrichment significance ($-\log_{10} P$ value) for genetic variants associated with diverse traits. Circles denote reference epigenome (column) of most significant enrichment for SNPs reported by a given study (row), defined by trait and publication (PubMed

human disease (see <http://www.roadmapepigenomics.org/publications>). This paper constitutes the first integrative analysis of all the reference epigenomes generated by the consortium, and represents an early component of the International Human Epigenome Consortium (<http://ihec-epigenomes.org/>), which seeks to extend such epigenomic maps to more than a thousand reference human epigenomes⁹⁴.

In this paper, we use this resource to gain insights into the epigenomic landscape, its dynamics across cell types, tissues and development, and its regulatory circuitry. We find that combinations of histone modification marks are highly informative of the methylation and accessibility levels of different genomic regions, while the converse is not always true. Genomic regions vary greatly in their association with active marks, with approximately 5% of each epigenome marked by enhancer or promoter signatures on average, which show increased association with expressed genes, and increased evolutionary conservation, while two-thirds of each reference epigenome on average are quiescent, and enriched in gene-poor and nuclear-lamina-associated stably repressed regions. Even though promoter and transcription associated marks are less dynamic than enhancer mark, each mark recovers biologically meaningful cell-type groupings when evaluated in relevant chromatin states, allowing a data-driven approach to learn relationships between

identifier, PMID). Tissue (Abbrev) and *P* value ($-\log_{10}$) of most significant enrichment are shown. Only rows and columns containing a value meeting a FDR of 2% are shown (see Extended Data Figs 11 and 12 for full matrix for all studies showing at least 2% FDR).

cell types, tissues and lineages. The coordinated activity patterns of enhancer regions enable us to cluster them into putative co-regulated modules, which are proximal to genes with common functions and phenotypes and enriched in regulatory motifs, enabling us to predict candidate upstream regulators.

We also demonstrate the usefulness of the resulting regulatory annotations for interpreting human genetic variation and disease. In an unbiased sampling across the GWAS catalogue, we find that genetic variants associated with complex traits are highly enriched in epigenomic annotations of trait-relevant tissues, providing insights on the likely relevant cell types underlying genome-wide significant loci. The GWAS enrichments in our analysis were strongest for enhancer-associated marks, consistent with their highly tissue-specific nature. However, promoter-associated and transcription-associated marks were also enriched, implicating several gene-regulatory levels as underlying genetic variants associated with complex traits. These results suggest that our data sets will be valuable in the study of human disease, as several companion papers explore in the context of autoimmune disorders^{95,96}, Alzheimer's disease^{91,97,98} and cancer^{99,100}.

Overall, our epigenomic data sets, regulatory annotations and integrative analyses have resulted in the most comprehensive map of the

human epigenomic landscape so far across the largest collection of primary cells and tissues. We expect that this map will be of broad use to the scientific and biomedical communities, for studies of genome interpretation, gene regulation, cellular differentiation, genome evolution, genetic variation and human disease.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 10 April 2014; accepted 21 January 2015.

1. Rivera, C. M. & Ren, B. Mapping human epigenomes. *Cell* **155**, 39–55 (2013).
2. Zhou, V. W., Goren, A. & Bernstein, B. E. Charting histone modifications and the functional organization of mammalian genomes. *Nature Rev. Genet.* **12**, 7–18 (2011).
3. Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Rev. Genet.* **13**, 484–492 (2012).
4. Smith, Z. D. & Meissner, A. DNA methylation: roles in mammalian development. *Nature Rev. Genet.* **14**, 204–220 (2013).
5. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Rev. Genet.* **10**, 57–63 (2009).
6. Park, P. J. ChIP-seq: advantages and challenges of a maturing technology. *Nature Rev. Genet.* **10**, 669–680 (2009).
7. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).
8. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5**, 621–628 (2008).
9. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).
10. Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genet.* **39**, 311–318 (2007).
11. Xie, W. *et al.* Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell* **153**, 1134–1148 (2013).
12. Zhu, J. *et al.* Genome-wide chromatin state transitions associated with developmental and environmental cues. *Cell* **152**, 642–654 (2013).
13. Nepf, S. *et al.* An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489**, 83–90 (2012).
14. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
15. Bernstein, B. E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nature Biotechnol.* **28**, 1045–1048 (2010).
16. Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553–560 (2007).
17. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007).
18. John, S. *et al.* Genome-scale mapping of DNase I hypersensitivity. *Curr. Protoc. Mol. Biol.* **Ch. 27**, Unit 21 27 (2013).
19. Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315–322 (2009).
20. Meissner, A. *et al.* Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res.* **33**, 5868–5877 (2005).
21. Weber, M. *et al.* Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nature Genet.* **37**, 853–862 (2005).
22. Maunakea, A. K. *et al.* Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature* **466**, 253–257 (2010).
23. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
24. Bernstein, B. E. *et al.* Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell* **120**, 169–181 (2005).
25. Bonasio, R., Tu, S. & Reinberg, D. Molecular signals of epigenetic states. *Science* **330**, 612–616 (2010).
26. Peters, A. H. *et al.* Partitioning and plasticity of repressive histone methylation states in mammalian chromatin. *Mol. Cell* **12**, 1577–1589 (2003).
27. Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108–112 (2009).
28. Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279–283 (2011).
29. Creighton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl Acad. Sci. USA* **107**, 21931–21936 (2010).
30. Cedar, H. & Bergman, Y. Linking DNA methylation and histone modification: patterns and paradigms. *Nature Rev. Genet.* **10**, 295–304 (2009).
31. Stevens, M. *et al.* Estimating absolute methylation levels at single-CpG resolution from methylation enrichment and restriction enzyme sequencing methods. *Genome Res.* **23**, 1541–1553 (2013).
32. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
33. Butterfield, Y. S. *et al.* JAGuar: Junction Alignments to Genome for RNA-Seq Reads. *PLoS ONE* **9**, e102398 (2014).
34. Coarfa, C. *et al.* Pash 3.0: A versatile software package for read mapping and integrative analysis of genomic and epigenomic variation using massively parallel DNA sequencing. *BMC Bioinform.* **11**, 572 (2010).
35. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
36. Fejes, A. P. *et al.* FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics* **24**, 1729–1730 (2008).
37. Landt, S. G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* **22**, 1813–1831 (2012).
38. Kunde-Ramamoorthy, G. *et al.* Comparison and quantitative verification of mapping algorithms for whole-genome bisulfite sequencing. *Nucleic Acids Res.* **42**, e43 (2014).
39. Harris, R. A. *et al.* Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nature Biotechnol.* **28**, 1097–1105 (2010).
40. Ernst, J. & Kellis, M. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nature Biotechnol.* <http://dx.doi.org/10.1038/nbt.3157> (in the press).
41. Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature Biotechnol.* **28**, 817–825 (2010).
42. Davydov, E. V. *et al.* Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLOS Comput. Biol.* **6**, e1001025 (2010).
43. Kagey, M. H. *et al.* Mediator and cohesin connect gene expression and chromatin architecture. *Nature* **467**, 430–435 (2010).
44. Stadler, M. B. *et al.* DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* **480**, 490–495 (2011).
45. Gascard, P. *et al.* Epigenetic and transcriptional determinants of the human breast. *Nature Commun.* <http://dx.doi.org/10.1038/ncomms7351> (in the press).
46. Mohn, F., Weber, M., Schubeler, D. & Roloff, T. C. Methylated DNA immunoprecipitation (MeDIP). *Methods Mol. Biol.* **507**, 55–64 (2009).
47. Elliott, G. *et al.* Intermediate DNA methylation is a conserved signature of genome regulation. *Nature Commun.* <http://dx.doi.org/10.1038/ncomms7363> (in the press).
48. Ji, H. *et al.* Comprehensive methylome map of lineage commitment from haematopoietic progenitors. *Nature* **467**, 338–342 (2010).
49. Meissner, A. *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454**, 766–770 (2008).
50. Gifford, C. A. *et al.* Transcriptional and epigenetic dynamics during specification of human embryonic stem cells. *Cell* **153**, 1149–1163 (2013).
51. Ziller, M. J. *et al.* Charting a dynamic DNA methylation landscape of the human genome. *Nature* **500**, 477–481 (2013).
52. Tsankov, A. M. *et al.* Transcription factor binding dynamics during human ESC differentiation. *Nature* <http://dx.doi.org/10.1038/nature14233> (this issue).
53. Ziller, M. J. *et al.* Dissecting neural differentiation regulatory networks through epigenetic footprinting. *Nature* <http://dx.doi.org/10.1038/nature13990> (this issue).
54. Xie, M. *et al.* DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape. *Nature Genet.* **45**, 836–841 (2013).
55. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnol.* **28**, 495–501 (2010).
56. Lowdon, R. F. *et al.* Regulatory network decoded from epigenomes of surface ectoderm-derived cell types. *Nat. Commun.* **5**, 5442 (2014).
57. Amin, V. *et al.* Epigenomic footprints across 111 reference epigenomes reveal tissue-specific epigenetic regulation of lincRNAs. *Nature Commun.* <http://dx.doi.org/10.1038/ncomms7370> (in the press).
58. Bernstein, B. E. *et al.* A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**, 315–326 (2006).
59. Hawkins, R. D. *et al.* Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell Stem Cell* **6**, 479–491 (2010).
60. Varley, K. E. *et al.* Dynamic DNA methylation across diverse human cell lines and tissues. *Genome Res.* **23**, 555–567 (2013).
61. Leung, D. *et al.* Integrative analysis of haplotype-resolved epigenomes across human tissues. *Nature* <http://dx.doi.org/10.1038/nature14217> (this issue).
62. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
63. Meuleman, W. *et al.* Constitutive nuclear lamina-genome interactions are highly conserved and associated with A/T-rich sequence. *Genome Res.* **23**, 270–280 (2013).
64. Guelen, L. *et al.* Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453**, 948–951 (2008).
65. Antequera, F., Boyes, J. & Bird, A. High levels of de novo methylation and altered chromatin structure at CpG islands in cell lines. *Cell* **62**, 503–514 (1990).
66. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.* **25**, 25–29 (2000).
67. Kohler, S. *et al.* The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res.* **42**, D966–D974 (2014).
68. Kheradpour, P. & Kellis, M. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res.* **42**, 2976–2987 (2014).
69. Hesselberth, J. R. *et al.* Global mapping of protein–DNA interactions *in vivo* by digital genomic footprinting. *Nature Methods* **6**, 283–289 (2009).

70. Kheradpour, P., Stark, A., Roy, S. & Kellis, M. Reliable prediction of regulator targets using 12 *Drosophila* genomes. *Genome Res.* **17**, 1919–1931 (2007).
71. Whitaker, J. W., Chen, Z. & Wang, W. Predicting the human epigenome from DNA motifs. *Nature Methods* <http://dx.doi.org/10.1038/nmeth.3065> (in the press).
72. Dixon, J. R. *et al.* Chromatin architecture reorganization during stem cell differentiation. *Nature* <http://dx.doi.org/10.1038/nature14222> (this issue).
73. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476–482 (2011).
74. Trynka, G. *et al.* Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nature Genet.* **45**, 124–130 (2013).
75. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001–D1006 (2014).
76. Franke, A. *et al.* Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature Genet.* **42**, 1118–1125 (2010).
77. Cooper, J. D. *et al.* Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. *Nature Genet.* **40**, 1399–1401 (2008).
78. Berndt, S. I. *et al.* Genome-wide association study identifies multiple risk loci for chronic lymphocytic leukemia. *Nature Genet.* **45**, 868–876 (2013).
79. Stahl, E. A. *et al.* Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nature Genet.* **42**, 508–514 (2010).
80. Barrett, J. C. *et al.* Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nature Genet.* **41**, 703–707 (2009).
81. Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119–124 (2012).
82. Yang, W. *et al.* Meta-analysis followed by replication identifies loci in or near *CDKN1B*, *TET3*, *CD80*, *DRAM1*, and *ARID5B* as associated with systemic lupus erythematosus in Asians. *Am. J. Hum. Genet.* **92**, 41–51 (2013).
83. Musunuru, K. *et al.* From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* **466**, 714–719 (2010).
84. Willy, P. J. *et al.* LXR, a nuclear receptor that defines a distinct retinoid response pathway. *Genes Dev.* **9**, 1033–1045 (1995).
85. Pasquali, L. *et al.* Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nature Genet.* **46**, 136–143 (2014).
86. Dalcik, H. *et al.* Expression of insulin-like growth factor in the placenta of intrauterine growth-retarded human fetuses. *Acta Histochem.* **103**, 195–207 (2001).
87. Lesch, K. P. *et al.* Molecular genetics of adult ADHD: converging evidence from genome-wide association and extended pedigree linkage studies. *J. Neural Transm.* **115**, 1573–1585 (2008).
88. Repunte-Canonigo, V. *et al.* A potential role for adiponectin receptor 2 (AdipoR2) in the regulation of alcohol intake. *Brain Res.* **1339**, 11–17 (2010).
89. Sawcer, S. *et al.* Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature* **476**, 214–219 (2011).
90. Heneka, M. T., Kummer, M. P. & Latz, E. Innate immune activation in neurodegenerative disease. *Nature Rev. Immunol.* **14**, 463–477 (2014).
91. Gjonneska, E. *et al.* Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease. *Nature* <http://dx.doi.org/10.1038/nature14252> (this issue).
92. Zhou, X. *et al.* Epigenomic annotation of genetic variants using the Roadmap Epigenome Browser. *Nature Biotechnol.* <http://dx.doi.org/10.1038/nbt.3158> (in the press).
93. Ward, L. D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* **40**, D930–D934 (2012).
94. Satterlee, J. S., Schubeler, D. & Ng, H. H. Tackling the epigenome: challenges and opportunities for collaboration. *Nature Biotechnol.* **28**, 1039–1044 (2010).
95. Farh, K. K. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* <http://dx.doi.org/10.1038/nature13835> (this issue).
96. Seumois, G. *et al.* Epigenomic analysis of primary human T cells reveals enhancers associated with TH2 memory cell differentiation and asthma susceptibility. *Nature Immunol.* **15**, 777–788 (2014).
97. De Jager, P. L. *et al.* Alzheimer's disease: early alterations in brain DNA methylation at ANK1, BIN1, RHBDL2 and other loci. *Nature Neurosci.* **17**, 1156–1163 (2014).
98. Lunnon, K. *et al.* Methylation profiling implicates cortical deregulation of ANK1 in Alzheimer's disease. *Nature Neurosci.* **17**, 1164–1170 (2014).
99. Polak, P. *et al.* Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* <http://dx.doi.org/10.1038/nature14221> (this issue).
100. Yao, L., Tak, Y. G., Berman, B. P. & Farnham, P. J. Functional annotation of colon cancer risk SNPs. *Nat. Commun.* **5**, 5114 (2014).
101. Zhou, X. *et al.* The Human Epigenome Browser at Washington University. *Nature Methods* **8**, 989–990 (2011).
102. Karolchik, D. *et al.* The UCSC Genome Browser Database. *Nucleic Acids Res.* **31**, 51–54 (2003).
103. Chadwick, L. H. The NIH Roadmap Epigenomics Program data resource. *Epigenomics* **4**, 317–324 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported by the NIH Common Fund as part of the NIH Roadmap Epigenomics Program through U01ES017155 (B.B. and A.M.), U01ES017154 (J.C. and M.M.), U01ES017166 (B.R.), U01ES017156 (J.S.), U01DA025956 (A.M. and A.B.), and by NHGRI through RC1HG005334, R01HG004037 and R01HG004037-S1 (M.K.), R01NS078839 (L.-H.T.), and by NIH ES017166, NSFC 91019016 and NBRPC 2012CB316503 (M.Q.Z.). Sample procurement was supported by grants 5R24HD000836 (I.A.G.) for staged fetal tissues;

P30AG10161, R01AG15819, R01AG17917 (D.A.B.) and U01AG46152 (P.L.D. and D.A.B.) for adult brain samples. This work was also supported by NIH fellowship grants F32HL110473 and K99HL119617 (S.L.), and NSF CAREER award 1254200 (J.E.). We acknowledge program leadership by members of the NIH Epigenomics Workgroup, especially J. S. Satterlee, F. L. Tyson, J. Rutter, K. A. McAllister, A. Haugen, C. Colvis (NCATS), J. Battey (NIDCD), L. Birnbaum (NIEHS) and N. Volkow (NIDA). We acknowledge feedback from our External Scientific Panel members M. Bartolomei, S. Baylin, S. Beck, A. Chakravarti, L. Jackson-Grusby, J. Lieb, S. Peckman, J. Quackenbush and S. Stice.

Author Contributions Details of author contributions are provided in the Roadmap Epigenomics Consortium list.

Author Information All data sets and analysis results are available at <http://compbio.mit.edu/roadmap/>. Browseable views of all data sets (as shown in Fig. 3) are available from the WashU Epigenome Browser¹⁰¹ at <http://epigenomegateway.wustl.edu/> and the UCSC Genome Browser¹⁰² at <http://genome.ucsc.edu/cgi-bin/hgTracks?db=hg19&hubUrl=http://vizhub.wustl.edu/VizHub/RoadmapReleaseAll.txt>. All primary data sets and protocols are available at REMC portal¹⁰³ at <http://www.roadmapepigenomics.org>, GEO data sets at <http://ncbi.nlm.nih.gov/geo/roadmap/epigenomics>, and the Human Epigenome Atlas at <http://epigenomeatlas.org>. Epigenomic annotations and motif predictions are incorporated into HaploReg for mining GWAS at <http://compbio.mit.edu/haploreg>. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.K. (manoli@mit.edu).



This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported licence. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons licence, users will need to obtain permission from the licence holder to reproduce the material. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-sa/3.0>

¹Computer Science and Artificial Intelligence Lab, Massachusetts Institute of Technology, 32 Vassar St, Cambridge, Massachusetts 02139, USA. ²The Broad Institute of Harvard and MIT, 415 Main Street, Cambridge, Massachusetts 02142, USA. ³Department of Genetics, Department of Computer Science, 300 Pasteur Dr., Lane Building, L301, Stanford, California 94305-5120, USA. ⁴Department of Biological Chemistry, University of California, Los Angeles, 615 Charles E Young Dr South, Los Angeles, California 90095, USA. ⁵Canada's Michael Smith Genome Sciences Centre, BC Cancer Agency, 675 West 10th Avenue, Vancouver, British Columbia V5Z 1L3, Canada. ⁶Department of Stem Cell and Regenerative Biology, 7 Divinity Ave, Cambridge, Massachusetts 02138, USA. ⁷Epigenome Center, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA. ⁸Department of Cellular and Molecular Medicine, Institute of Genomic Medicine, Moores Cancer Center, Department of Chemistry and Biochemistry, University of California San Diego, 9500 Gilman Drive, La Jolla, California 92093, USA. ⁹Genomic Analysis Laboratory, Howard Hughes Medical Institute & The Salk Institute for Biological Studies, 10010 N. Torrey Pines Road, La Jolla, California 92037, USA. ¹⁰Department of Genome Sciences, University of Washington, 3720 15th Ave. NE, Seattle, Washington 98195, USA. ¹¹Biology Department, Massachusetts Institute of Technology, 31 Ames St, Cambridge, Massachusetts 02142, USA. ¹²The Picower Institute for Learning and Memory, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, 43 Vassar St, Cambridge, Massachusetts 02139, USA. ¹³Ludwig Institute for Cancer Research, 9500 Gilman Drive, La Jolla, California 92093, USA. ¹⁴Department of Neurosurgery, Helen Diller Family Comprehensive Cancer Center, University of California San Francisco, 1450 3rd Street, San Francisco, California 94158, USA. ¹⁵Department of Pathology, University of California San Francisco, 513 Parnassus Avenue, San Francisco, California 94143-0511, USA. ¹⁶Department of Medicine, Division of Medical Genetics, University of Washington, 2211 Elliot Avenue, Seattle, Washington 98121, USA. ¹⁷Department of Computer Science & Engineering, University of Connecticut, 371 Fairfield Way, Storrs, Connecticut 06269, USA. ¹⁸Department of Microbiology and Immunology and Centre for High-Throughput Biology, University of British Columbia, 2125 East Mall, Vancouver, British Columbia V6T 1Z4, Canada. ¹⁹Bioinformatics Group, Department of Molecular Biology, Division of Biology, Faculty of Science, University of Zagreb, Horvatovac 102a, 10000 Zagreb, Croatia. ²⁰Department of Molecular and Cell Biology, Center for Systems Biology, The University of Texas, Dallas, NSERL, RL10, 800 W Campbell Road, Richardson, Texas 75080, USA. ²¹Department of Genetics, Center for Genome Sciences and Systems Biology, Washington University in St Louis, 4444 Forest Park Ave, St Louis, Missouri 63108, USA. ²²Institute for Molecular Bioscience, University of Queensland, St Lucia, Queensland 4072, Australia. ²³Brigham & Women's Hospital, 75 Francis Street, Boston, Massachusetts 02115, USA. ²⁴Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, New York 11794-3600, USA. ²⁵Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA. ²⁶Molecular and Human Genetics Department, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA. ²⁷Harvard Medical School, 25 Shattuck St, Boston, Massachusetts 02115, USA. ²⁸Department of Biochemistry, Keck School of Medicine, University of Southern California, 1450 Biggy Street, Los Angeles, California 90089-9601, USA. ²⁹ObGyn, Reproductive Sciences, University of California San Francisco, 35 Medical Center Way, San Francisco, California 94143, USA. ³⁰Center for

Biomolecular Sciences and Engineering, University of Santa Cruz, 1156 High Street, Santa Cruz, California 95064, USA. ³¹Department of Molecular Biology and Biochemistry, Simon Fraser University, 8888 University Drive, Burnaby, British Columbia V5A 1S6, Canada. ³²Department of Medical Genetics, University of British Columbia, 2329 West Mall, Vancouver, BC, Canada, V6T 1Z4. ³³Dan L. Duncan Cancer Center, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA. ³⁴Department of Microbiology and Immunology, Diabetes Center, University of California, San Francisco, 513 Parnassus Ave, San Francisco, California 94143-0534, USA. ³⁵University of Wisconsin, Madison, Wisconsin 53715, USA. ³⁶USDA/ARS Children's Nutrition Research Center, Baylor College of Medicine, 1100 Bates Street, Houston, Texas 77030, USA. ³⁷Bioinformatics Division, Center for Synthetic and Systems Biology, TNLST, Tsinghua University, Beijing 100084, China. ³⁸National Institute of Environmental Health Sciences, 111 T.W. Alexander Drive, Research Triangle Park, North Carolina 27709, USA. ³⁹Massachusetts General Hospital, 55 Fruit St, Boston, Massachusetts 02114, USA. ⁴⁰Howard Hughes Medical Institute, 4000 Jones Bridge Road, Chevy Chase, Maryland 20815-6789, USA. ⁴¹Morgridge Institute for Research, 330 N. Orchard Street, Madison, Wisconsin 53707, USA. ⁴²Department of Computer Science and Engineering, Washington University in St. Louis, St. Louis, Missouri 63130, USA.

Roadmap Epigenomics Consortium

Integrative analysis coordination Anshul Kundaje^{1,2,3}, Wouter Meuleman^{1,2}, Jason Ernst^{1,2,4}, Misha Bilenky⁵; **Integrative analysis leads (equal contributors)** Angela Yen^{1,2}, Alireza Heravi-Moussavi⁶, Pouya Kheradpour^{1,2}, Zhizhou Zhang^{1,2}, Jianrong Wang^{1,2}, Michael J. Ziller^{2,6}, Viren Amin⁷, John W. Whitaker⁸, Matthew D. Schultz⁹, Lucas D. Ward^{1,2}, Abhishek Sarkar^{1,2}, Gerald Quon^{1,2}, Richard S. Sandstrom¹⁰, Matthew L. Eaton^{1,2}, Yi-Chieh Wu^{1,2}, Andreas R. Pfennig^{1,2}, Xincheng Wang^{1,2,11}, Melina Claussnitzer^{1,2}, Yaping Liu^{1,2}; **Data production and processing leads (equal contributors)** Cristian Coarfa⁷, R. Alan Harris⁷, Noam Shores⁷, Charles B. Epstein⁷, Elizabetha Gjonneska^{2,12}, Danny Leung^{8,13}, Wei Xie^{8,13}, R. David Hawkins^{8,13}, Ryan Lister⁹, Chibo Hong¹⁴, Philippe Gascard¹⁵, Andrew J. Mungall⁵, Richard Moore⁵, Eric Chuah⁵, Angela Tam⁵, Theresa K. Canfield¹⁰, R. Scott Hansen¹⁶, Rajinder Kaul¹⁶, Peter J. Sabo¹⁰; **Integrative analysis co-leads** Mukul S. Bansal^{1,2,17}, Annaick Carles¹⁸, Jesse R. Dixon^{8,13}, Kai-How Farh², Soheil Feizi^{1,2}, Rosa Karlic¹⁹, Ah-Ram Kim^{1,2}, Ashwinikumar Kulkarni²⁰, Daofeng Li²¹, Rebecca Lowdon²¹, GiNelli Elliott²¹, Tim R. Mercer²², Shane J. Neph¹⁰, Vitor Onuchic⁷, Paz Polak^{2,23}, Nisha Rajagopal^{8,13}, Pradipta Ray²⁰, Richard C. Sallari^{1,2}, Kyle T. Siebenthal¹⁰, Nicholas A. Sinnott-Armstrong^{1,2}, Michael Stevens^{21,58}, Robert E. Thurman¹⁰, Jie Wu^{24,25}, Bo Zhang²², Xin Zhou²¹; **Analysis and production contributors** Nezar Abdenur^{1,2}, Mazhar Adil^{26,27}, Martin Akerman²⁵, Luis Barrera^{1,2}, Jessica Antosiewicz-Bourget²⁸, Tracy Ballinger²⁹, Michael J. Barnes¹⁵, Daniel Bates¹⁰, Robert J. A. Bell¹⁴, David A. Bennett³⁰, Katherine Bianco³¹, Christoph Bock², Patrick Boyle², Jan Brinchmann³², Pedro Caballero-Campo³³, Raymond Camahort³⁴, Marlene J. Carrasco-Alfonso³⁴, Timothy Charnecki⁷, Huaming Chen⁹, Zhao Chen⁸, Jeffrey B. Cheng⁵⁴, Stephanie Cho⁵, Andy Chu⁵, Wen-Yu Chung²⁰, Chad Cowan³⁴, Qixia Athena Deng⁵, Vikram Deshpande²⁶, Morgan Diegel¹⁰, Bo Ding⁸, Timothy Durham², Lorigail Echipe⁵⁵, Lee Edsall¹³, David Flowers³⁷, Olga Genbacev-Krtolica³¹, Casey Gifford², Shawn Gillespie²⁶, Erika Giste¹⁰, Ian A. Glass³⁸, Andreas Gnirke², Matthew Gormley³¹, Honggang Gu², Junchen Gu²¹, David A. Hafler³⁹, Matthew J. Hangauer⁴⁰, Manoj Hariharan⁹, Meital Hatan², Eric Haugen¹⁰, Yungpe He³⁷, Shelly Heimfeld³⁷, Sarah Herlofson³², Zhonggang Hou²⁸, Richard Humbert¹⁰, Robbyn Issner², Andrew R. Jackson⁷, Haiyang Jia⁸, Peng Jiang²⁸, Audra K. Johnson¹⁰, Theresa Kadlec^{41,42}, Baljit Kamoh⁵, Mirhan Kapidzic³¹, Jim Kent²⁹, Audrey Kim^{8,13}, Markus Kleinewietfeld³⁹, Sarit Klugman³¹, Jayanthi Krishnan^{1,2}, Samantha Kuan¹³, Tanya Kutuyian¹⁰, Ah-Young Lee¹³, Kristen Lee¹⁰, Jian Li⁷, Nan Li⁸, Yan Li⁸, Keith L. Ligon⁴³, Shin Lin⁹, Yiling Lin⁹, Jie Liu⁸, Yuxuan Liu²⁰, C. John Luckey³⁴, Yussanne P. Ma², Cecile Maire⁴³, Alexander Marson³⁵, John S. Mattick^{44,45}, Michael Mayo⁵, Michael McMaster³¹, Hayden Metsky^{1,2}, Tarjei Mikkelsen², Diane Miller⁵, Mohammad Miri²⁶, Eran Mukamel⁹, Raman P. Nagarajan¹⁴, Fidencio Neri¹⁰, Joseph Nery⁹, Tung Nguyen⁵, Henriette O'Geen⁵⁵, Sameer Patilthakar⁷, Thalia Papayannopoulou¹⁶, Mattia Pelizzola⁹, Patrick Plettner⁵, Nicholas E. Propson²⁸, Sriram Raghuraman⁷, Brian J. Raney²⁹, Anthony Raubitschek⁴⁶, Alex P. Reynolds¹⁰, Hunter Richards⁴⁰, Kevin Riehle⁷, Paolo Rinaldo³³, Joshua F. Robinson³¹, Nicole B. Rockweiler²¹, Evan Rosen³⁴, Eric Rynes¹⁰, Jacqueline Schein⁹, Renee Sears²¹, Terrence Sejnowski⁹, Anthony Shafer¹⁰, Li Shen^{8,56}, Robert Shoemaker⁸, Mahvash Sigaroudinia¹⁵, Igor Slukvin⁵⁷, Sandra Stehling-Sun¹⁰, Ron Stewart²⁸, Sailakshmi Subramanian⁷, Kran Suknuntha²⁸, Scott Swanson²⁸, Shulan Tian⁵⁷, Hannah Tilden³¹, Linus Tsai³⁴, Mark Ulrich⁹, Ian Vaughn⁴⁰, Jeff Vierstra¹⁰, Shinny Vong¹⁰, Ulrich Wagner¹³, Hao Wang¹⁰, Tao Wang⁵, Yunfei Wang²⁰, Arthur Weiss⁴¹, Holly Whitton², Andre Wildberg⁸, Heather Witt³⁶, Kyoung-Jae Won⁸, Mingchao Xie²¹, Xiaoyun Xing²¹, Iris Xu^{1,2}, Zhenyu Xuan²⁰, Zhen Ye¹³, Chia-an Yen¹³, Pengzhi Yu²⁸, Xian Zhang⁸, Xiaolan Zhang², Jianxin Zhao¹⁵, Yan Zhou³¹, Jiang Zhu²⁶, Yun Zhu⁸, Steven Ziegler⁴⁶; **Co-principal investigators** Arthur E. Beaudet⁴⁷, Laurie A. Boyer¹¹, Philip L. De Jager^{2,23,34}, Peggy J. Farnham³⁶, Susan J. Fisher³¹, David Haussler²⁹, Steven J. M. Jones^{5,48,49}, Wei Li⁵⁰, Marco A. Marra^{5,49}, Michael T. Manus⁴⁰, Shamir Sunyaev^{2,23,34}, James A. Thomson^{28,57}, Thea D. Tlsty¹⁵, Li-Huei Tsai^{2,12}, Wei Wang⁸, Robert A. Waterland⁵¹, Michael Q. Zhang^{20,52}; **Scientific program management** Lisa H. Chadwick⁵³; **Principal investigators** Bradley E. Bernstein^{2,26,42}, Joseph F. Costello¹⁴, Joseph R. Ecker⁹, Martin Hirst^{5,18}, Alexander Meissner^{2,6}, Aleksandar Milosavljevic⁷, Bing Ren^{8,13}, John A. Stamatoyannopoulos¹⁰, Ting Wang²¹ & Manolis Kellis^{1,2}

¹Computer Science and Artificial Intelligence Lab, Massachusetts Institute of Technology, 32 Vassar St, Cambridge, Massachusetts 02139, USA. ²The Broad Institute of Harvard

and MIT, 415 Main Street, Cambridge, Massachusetts 02142, USA. ³Department of Genetics, Department of Computer Science, 300 Pasteur Dr., Lane Building, L301, Stanford, California 94305-5120, USA. ⁴Department of Biological Chemistry, University of California, Los Angeles, 615 Charles E Young Dr South, Los Angeles, California 90095, USA. ⁵Canada's Michael Smith Genome Sciences Centre, BC Cancer Agency, 675 West 10th Avenue, Vancouver, British Columbia V5Z 1L3, Canada. ⁶Department of Stem Cell and Regenerative Biology, 7 Divinity Ave, Cambridge, Massachusetts 02138, USA. ⁷Epigenome Center, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA. ⁸Department of Cellular and Molecular Medicine, Institute of Genomic Medicine, Moores Cancer Center, Department of Chemistry and Biochemistry, University of California San Diego, 9500 Gilman Drive, La Jolla, California 92093, USA. ⁹Genomic Analysis Laboratory, Howard Hughes Medical Institute & The Salk Institute for Biological Studies, 10010 N. Torrey Pines Road, La Jolla, California 92037, USA. ¹⁰Department of Genome Sciences, University of Washington, 3720 15th Ave. NE, Seattle, Washington 98195, USA. ¹¹Biology Department, Massachusetts Institute of Technology, 31 Ames St, Cambridge, Massachusetts 02142, USA. ¹²The Picower Institute for Learning and Memory, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, 43 Vassar St, Cambridge, Massachusetts 02139, USA. ¹³Ludwig Institute for Cancer Research, 9500 Gilman Drive, La Jolla, California 92093, USA. ¹⁴Department of Neurosurgery, Helen Diller Family Comprehensive Cancer Center, University of California San Francisco, 1450 3rd Street, San Francisco, California 94158, USA. ¹⁵Department of Pathology, University of California San Francisco, 513 Parnassus Avenue, San Francisco, California 94143-0511, USA. ¹⁶Department of Medicine, Division of Medical Genetics, University of Washington, 2211 Elliot Avenue, Seattle, Washington 98121, USA. ¹⁷Department of Computer Science & Engineering, University of Connecticut, 371 Fairfield Way, Storrs, Connecticut 06269, USA. ¹⁸Department of Microbiology and Immunology and Centre for High-Throughput Biology, University of British Columbia, 2125 East Mall, Vancouver, British Columbia V6T 1Z4, Canada. ¹⁹Bioinformatics Group, Department of Molecular Biology, Division of Biology, Faculty of Science, University of Zagreb, Horvatova 102a, 10000 Zagreb, Croatia. ²⁰Department of Molecular and Cell Biology, Center for Systems Biology, The University of Texas, Dallas, NSERL, RL10, 800 W Campbell Road, Richardson, Texas 75080, USA. ²¹Department of Genetics, Center for Genome Sciences and Systems Biology, Washington University in St. Louis, 4444 Forest Park Ave, St. Louis, Missouri 63108, USA. ²²Institute for Molecular Bioscience, University of Queensland, St Lucia, Queensland 4072, Australia. ²³Brigham & Women's Hospital, 75 Francis Street, Boston, Massachusetts 02115, USA. ²⁴Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, New York 11794-3600, USA. ²⁵Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA. ²⁶Massachusetts General Hospital, 55 Fruit St, Boston, Massachusetts 02114, USA. ²⁷University of Virginia, School of Medicine, 1340 Jefferson Park Ave, Charlottesville, Virginia 22908, USA. ²⁸Morgridge Institute for Research, 330 N. Orchard Street, Madison, Wisconsin 53707, USA. ²⁹Center for Biomolecular Sciences and Engineering, University of Santa Cruz, 1156 High Street, Santa Cruz, California 95064, USA. ³⁰Rush University Medical Center, 1653 W Congress Pkwy, Chicago, Illinois 60612, USA. ³¹ObGyn, Reproductive Sciences, University of California San Francisco, 35 Medical Center Way, San Francisco, California 94143, USA. ³²Rikshospitalet University Hospital, Sognsvannsveien 20, 0372 Oslo, Norway. ³³Reproductive Endocrinology and Infertility, University of California San Francisco, 2356 Sutter St, San Francisco, California 94115, USA. ³⁴Harvard Medical School, 25 Shattuck St, Boston, Massachusetts 02115, USA. ³⁵UCSF School of Medicine, 513 Parnassus Avenue, San Francisco, California 94143, USA. ³⁶Department of Biochemistry, Keck School of Medicine, University of Southern California, 1450 Biggy Street, Los Angeles, California 90089-9601, USA. ³⁷Fred Hutchinson Cancer Research Center, 1100 Fairview Ave N, Seattle, Washington 98109, USA. ³⁸Department of Pediatrics, Seattle Children's Hospital/University of Washington, 4800 Sand Point Way NE, Seattle, Washington 98105, USA. ³⁹Yale School of Medicine, 333 Cedar Street, New Haven, Connecticut 06510, USA. ⁴⁰Department of Microbiology and Immunology, Diabetes Center, University of California, San Francisco, 513 Parnassus Ave, San Francisco, California 94143-0534, USA. ⁴¹School of Medicine, University of California San Francisco, 513 Parnassus Avenue, San Francisco, California 94143, USA. ⁴²Howard Hughes Medical Institute, 4000 Jones Bridge Road, Chevy Chase, Maryland 20815-6789, USA. ⁴³Center for Molecular Oncologic Pathology, Dana-Farber Cancer Institute/Brigham and Women's Hospital, 450 Brookline Avenue, Boston, Massachusetts 02215, USA. ⁴⁴Garvan Institute of Medical Research, 384 Victoria Street, Darlinghurst NSW 2010, Australia. ⁴⁵St Vincent's Clinical School, University of New South Wales, Sydney, New South Wales 2052, Australia. ⁴⁶Immunology Research Program, Benaroya Research Institute, 1201 Ninth Avenue, Seattle, Washington 98101, USA. ⁴⁷Molecular and Human Genetics Department, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA. ⁴⁸Department of Molecular Biology and Biochemistry, Simon Fraser University, 8888 University Drive, Burnaby, British Columbia V5A 1S6, Canada. ⁴⁹Department of Medical Genetics, University of British Columbia, 2329 West Mall, Vancouver, BC, Canada, V6T 1Z4. ⁵⁰Dan L. Duncan Cancer Center, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA. ⁵¹USDA/ARS Children's Nutrition Research Center, Baylor College of Medicine, 1100 Bates Street, Houston, Texas 77030, USA. ⁵²Bioinformatics Division, Center for Synthetic and Systems Biology, TNLST, Tsinghua University, Beijing 100084, China. ⁵³National Institute of Environmental Health Sciences, 111 T.W. Alexander Drive, Research Triangle Park, North Carolina 27709, USA. ⁵⁴Department of Dermatology, University of California San Francisco, 1701 Divisadero Street, San Francisco, California 94143, USA. ⁵⁵UC Davis Genome Center, 451 Health Sciences Drive, Davis, California 95616, USA. ⁵⁶Department of Neurosciences, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA. ⁵⁷University of Wisconsin, Madison, Wisconsin 53715, USA. ⁵⁸Department of Computer Science and Engineering, Washington University in St. Louis, St. Louis, Missouri 63130, USA.

METHODS

No statistical methods were used to predetermine sample size.

Data matrix, primary analysis and processing quality control. All genome-wide maps of histone modifications, DNA accessibility, DNA methylation and RNA expression are freely available online. Links for raw sequencing data deposited at the Short Read Archive or dbGAP are available at <http://www.ncbi.nlm.nih.gov/geo/roadmap/epigenomics/>. All primary processed data (including mapped reads) for profiling experiments are contained within Release 9 of the Human Epigenome Atlas (<http://www.epigenomeatlas.org>). Complete metadata associated with each data set in this collection is archived at GEO and describes samples, assays, data processing details and quality metrics collected for each profiling experiment.

Release 9 of the compendium contains uniformly pre-processed and mapped data from multiple profiling experiments (technical and biological replicates from multiple individuals and/or data sets from multiple centres). To reduce redundancy, improve data quality and achieve uniformity required for our integrative analyses, experiments were subjected to additional processing to obtain comprehensive data for 111 consolidated epigenomes (see sections below for additional details). Numeric epigenome identifiers (EIDs; for example, E001) and mnemonics for epigenome names were assigned for each of the consolidated epigenomes. Supplementary Table 1 (QCSummary sheet) summarizes the mapping of the individual Release 9 samples to the consolidated epigenome IDs. Key metadata such as age, sex, anatomy, epigenome class (see Supplementary Table 1, EpigenomeClassSummary sheet), ethnicity and solid/liquid status were summarized for the consolidated epigenomes. Data sets corresponding to 16 cell lines from the ENCODE project (with epigenome IDs ranging from E114 to E129) were also used in the integrative analyses²³. All data sets from the 127 consolidated epigenomes were subjected to processing filters to ensure uniformity in terms of read-length-based mappability and sequencing depth as described below.

Each of the 127 epigenomes included consolidated ChIP-seq data sets for a core set of histone modifications—H3K4me1, H3K4me3, H3K27me3, H3K36me3, H3K9me3—as well as a corresponding whole-cell extract sequenced control. Ninety-eight epigenomes and sixty-two epigenomes had consolidated H3K27ac and H3K9ac histone ChIP-seq data sets, respectively. A smaller subset of epigenomes had ChIP-seq data sets for additional histone marks, giving a total of 1,319 consolidated data sets (Supplementary Table 1, QCSummary sheet). 53 epigenomes had DNA accessibility (DNase-seq) data sets. Fifty-six epigenomes had mRNA-seq gene expression data. For the 127 consolidated epigenomes, a total of 104 DNA methylation data sets across 95 epigenomes involved either bisulfite treatment (WGBS or RRBS assays) or a combination of MeDIP-seq and MRE-seq assays. In addition to the 1,936 data sets analysed here across 111 reference epigenomes, the NIH Roadmap Epigenomics Project has generated an additional 869 genome-wide data sets, linked from GEO, the Human Epigenome Atlas, and NCBI, and also publicly and freely available.

RNA-seq uniform processing and quantification for consolidated epigenomes. We uniformly reprocessed mRNA-seq data sets from 56 reference epigenomes that had RNA-seq data. For RNA-seq analysis, after library construction⁴⁵, we aligned 75-bp-long or 100-bp-long reads using the BWA aligner, and generated read coverage profiles separately for positive and negative strand strand-specific libraries. We used several QC metrics for the RNA-seq library, including intron–exon ratio, intergenic reads fraction, strand specificity (for stranded RNA-seq protocols), 3'–5' bias, GC bias and RPKM discovery rate (Supplementary Table 1, RNaseqQCSummary sheet). We quantified exon and gene expression using a modified RPKM measure⁸, whereby we used the total number of reads aligned into coding exons for the normalization factor in RPKM calculations, and excluded reads from the mitochondrial genome, reads falling into genes coding for ribosomal proteins, and reads falling into top 0.5% expressed exons. RPKM for a gene was calculated using the total number of reads aligned into all merged exons for a gene normalized by total exonic length. The resulting files contain RPKM values for all annotated exons and coding and non-coding genes (excluding ribosomal genes), as well as introns (Gencode V10 annotations were used). We also report the coordinates of all significant intergenic RNA-seq contigs not overlapping the annotated genes.

ChIP-seq and DNase-seq uniform reprocessing for consolidated epigenomes. *Read mapping.* Sequenced data sets from the Release 9 of the Epigenome Atlas involved mapping a total of 150.21 billion sequencing reads onto hg19 assembly of the human genome using the PASH read mapper³⁴. These read mappings were used (except for RNA-seq data sets, which were mapped as described above) for constructing the 111 consolidated epigenomes. Only uniquely mapping reads were retained and multiply-mapping reads were filtered out. BED files containing the mapped reads were obtained from <http://www.epigenomeatlas.org>. Alignment parameters for each assay type and experiment are specified in the associated publicly accessible Release 9 metadata archived at GEO. For the ENCODE data sets, BAM files containing mapped reads were downloaded from <http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/>. Only uniquely mapping reads were retained and multiply mapping reads were discarded.

Mappability filtering, pooling and subsampling. The raw Release 9 read alignment files contain reads that are pre-extended to 200 bp. However, there were significant differences in the original read lengths across the Release 9 raw data sets reflecting differences between centres and changes of sequencing technology during the course of the project (36 bp, 50 bp, 76 bp and 100 bp). To avoid artificial differences due to mappability, for each consolidated data set the raw mapped reads were uniformly truncated to 36 bp and then refiltered using a 36-bp custom mappability track to only retain reads that map to positions (taking strand into account) at which the corresponding 36-mers starting at those positions are unique in the genome. Filtered data sets were then merged across technical/biological replicates, and where necessary to obtain a single consolidated sample for every histone mark or DNase-seq in each standardized epigenome. Supplementary Table 1 summarizes the mapping of the individual Release 9 primary data sample files to the consolidated data files corresponding to the 127 consolidated reference epigenomes.

To avoid artificial differences in signal strength due to differences in sequencing depth, all consolidated histone mark data sets (except the additional histone marks the seven deeply profiled epigenomes, Fig. 2j) were uniformly subsampled to a maximum depth of 30 million reads (the median read depth over all consolidated samples). For the seven deeply profiled reference epigenomes (Fig. 2j), histone mark data sets were subsampled to a maximum of 45 million reads (median depth). The consolidated DNase-seq data sets were subsampled to a maximum depth of 50 million reads (median depth). These uniformly subsampled data sets were then used for all further processing steps (peak calling, signal coverage tracks, chromatin states).

Peak calling. For the histone ChIP-seq data, the MACSv2.0.10 peak caller was used to compare ChIP-seq signal to a corresponding whole-cell extract (WCE) sequenced control to identify narrow regions of enrichment (peaks) that pass a Poisson *P* value threshold 0.01, broad domains that pass a broad-peak Poisson *P* value of 0.1 and gapped peaks which are broad domains ($P < 0.1$) that include at least one narrow peak ($P < 0.01$) (<https://github.com/taoliu/MACS/>)³². Fragment lengths for each data set were pre-estimated using strand cross-correlation analysis and the SPP peak caller package (<https://code.google.com/p/phantompeakqualtools/>)³⁷ and these fragment length estimates were explicitly used as parameters in the MACS2 program ($-\text{shift-size} = \text{fragment_length}/2$).

For DNase-seq data, we used two methods to identify DNase I accessible sites. First, the Hotspot algorithm was used to identify fixed-size (150 bp) DNase hypersensitive sites, and more general-sized regions of DNA accessibility (hotspots) using an FDR of 0.01 (<http://www.uwencode.org/proj/hotspot/>)¹⁰⁴. MACSv2.0.10 was also used to call narrow peaks using the same settings specified above for the histone mark narrow peak calling.

Narrow peaks and broad domains were also generated for the unconsolidated, 36-bp mappability filtered histone mark ChIP-seq and DNase-seq Release 9 data sets using MACSv2.0.10 with the same settings as specified above.

Genome-wide signal coverage tracks. We used the signal processing engine of the MACSv2.0.10 peak caller to generate genome-wide signal coverage tracks. Whole-cell extract was used as a control for signal normalization for the histone ChIP-seq coverage. Each DNase-seq data set was normalized using simulated background data sets generated by uniformly distributing equivalent number of reads across the mappable genome. We generated two types of tracks that use different statistics based on a Poisson background model to represent per-base signal scores. Briefly, reads are extended in the 5' to 3' direction by the estimated fragment length. At each base, the observed counts of ChIP-seq/DNase I-seq extended reads overlapping the base are compared to corresponding dynamic expected background counts (λ_{local}) estimated from the control data set. λ_{local} is defined as $\max(\lambda_{\text{BG}}, \lambda_{1k}, \lambda_{5k}, \lambda_{10k})$ where λ_{BG} is the expected counts per base assuming a uniform distribution of control reads across all mappable bases in the genome and $\lambda_{1k}, \lambda_{5k}, \lambda_{10k}$ are expected counts estimated from the 1 kb, 5 kb and 10 kb window centred at the base. λ_{local} is adjusted for the ratio of the sequencing depth of ChIP-seq/DNase-seq data set relative to the control data set. The two types of signal score statistics computed per base are as follows.

(1) Fold-enrichment ratio of ChIP-seq or DNase counts relative to expected background counts λ_{local} . These scores provide a direct measure of the effect size of enrichment at any base in the genome.

(2) Negative \log_{10} of the Poisson *P*-value of ChIP-seq or DNase counts relative to expected background counts λ_{local} . These signal confidence scores provide a measure of statistical significance of the observed enrichment.

The $-\log_{10}(P \text{ value})$ scores provide a convenient way to threshold signal (for example, 2 corresponds to a *P* value threshold of 1×10^{-2}), similar to what is used in identifying enriched regions (peak calling). We recommend using the signal confidence score tracks for visualization. A universal threshold of 2 provides good separation between signal and noise. Both types of signal tracks were also generated for the unconsolidated data sets using the same parameter settings described above.

Quality control. For the primary Release 9 data sets, data quality enrichment scores were computed as the fraction of the uniquely mapped reads overlapping with areas of enrichment. Several methods were employed to select signal enrichment regions. The SPOT quality score was computed based on regions identified with the HotSpot peak caller¹⁰⁴; the FindPeaks quality score was inferred based on peak calls made using the FindPeaks³⁶ software; finally, a Poisson metric was derived by modelling the read distribution in genome-tiling 1,000-bp windows with a Poisson distribution and selecting as enriched regions windows with $P < 0.05$. All the quality scores in Release 9 are in agreement, with strong pairwise correlation (Pearson correlation > 0.9). Concordance between centres was confirmed and data analysis pipeline was validated at the outset of the project using data sets for the H1 cell line. The same pipeline was subsequently used to produce Release 9 data. ChIP-seq data for six histone modifications (H3K4me3, H3K27me3, H3K9ac, H3K9me3, H3K36me3 and H3K4me1) were independently generated for the H1 cell line by three REMCs (Broad, UCSD, UCSF-UBC). To quantify concordance, the reads from each experiment were mapped (Level 1 data), read density tracks (Level 2 data) were generated using the EDACC's primary data processing pipeline, and finally Pearson correlation coefficients were computed between each pair of experiments, as well as between experiments and H1 input acting as a control for background correlation between signals (Supplementary Table 2). The methylome processing pipeline was characterized experimentally on four independent samples^{38,39}.

For the uniformly reprocessed and consolidated ChIP-seq and DNase-seq data sets, strand cross-correlation measures were used to estimate signal-to-noise ratios (<https://code.google.com/p/phantompeakqualtools/>)³⁷. Data sets for each mark were rank-ordered based on the normalized strand cross-correlation coefficient (NSC) and flagged if the scores were significantly below the median value or in the range of NSC values for WCE extract controls. Consolidated data sets with extremely low sequencing depth (< 10 M reads) were also flagged. Each standardized epigenome was then manually assigned a subjective quality flag of 1 (high), 0 (medium) or -1 (low), based on the number of flagged data sets it contained. The SPOT, FindPeaks and Poisson quality scores were also recomputed for the consolidated data sets. We observed high correlations of the NSC scores with the SPOT (Pearson correlation of 0.7) and FindPeaks scores (Pearson correlation of 0.65). All QC measures are provided in Supplementary Table 1 (Sheets QCSummary and AdditionalQCScores).

To identify potential antibody cross-reactivity or mislabelling issues, a pairwise correlation heat map (Extended Data Fig. 1e) was computed across all consolidated data sets for H3K4me1, H3K4me3, H3K36me3, H3K27me3, H3K9me3, H3K27ac, H3K9ac and DNase. We computed the Pearson correlation between all pairs of the signal tracks based on signal in chromosomes 1–22 and chromosome X. We used the signal confidence score tracks ($-\log_{10}(\text{Poisson } P \text{ value})$) where we first computed the average signal scores within each consecutive 25-bp interval. To order the experiments in the heat map we defined the distance between two pairs of experiments as 1-correlation value and used a travelling salesman problem formulation¹⁰⁵. **Methylation data cross-assay standardization and uniform processing for consolidated epigenomes.** We used PASH³⁸ alignments for the WGBS and RRBS read alignments. From the number of converted and unconverted reads at each individual CpG the total coverage and fractional methylation were reported. The data were uniformly post-processed and formatted into two matrices for each chromosome. One matrix contained read coverage information for each base (C and G) in every CpG (row) and for each reference epigenome (column). Another matrix similarly contained fractional methylation ranging from 0 to 1. For the locations where coverage was ≤ 3 we considered data as missing. For MeDIP/MRE methylation data we used the output of the mCRF tool³¹ that reports fractional methylation in the range from 0 to 1 and uses an internal BWA mapping. The mCRF results were combined in a single matrix per chromosome for all reference epigenomes where available.

Chromatin state learning. To capture the significant combinatorial interactions between different chromatin marks in their spatial context (chromatin states) across 127 epigenomes, we used ChromHMM v.1.10¹⁰⁶, which is based on a multivariate Hidden Markov Model.

'Core' 15-state model. A ChromHMM model applicable to all 127 epigenomes was learned by virtually concatenating consolidated data corresponding to the core set of five chromatin marks assayed in all epigenomes (H3K4me3, H3K4me1, H3K36me3, H3K27me3, H3K9me3). The model was trained on 60 epigenomes with highest-quality data (Fig. 2k), which provided sufficient coverage of the different lineages and tissue types (Supplementary Table 1; Sheet QCSummary). The ChromHMM parameters used were as follows: reads were shifted in the 5' to 3' direction by 100 bp. For each consolidated ChIP-seq data set, read counts were computed in non-overlapping 200-bp bins across the entire genome. Each bin was discretized into two levels, 1 indicating enrichment and 0 indicating no enrichment. The binarization was performed by comparing ChIP-seq read counts to corresponding whole-cell extract control read counts within each bin and using a Poisson

P value threshold of 1×10^{-4} (the default discretization threshold in ChromHMM). We trained several models in parallel mode with the number of states ranging from 10 states to 25 states. We decided to use a 15-state model (Fig. 4a–f and Extended Data Fig. 2b) for all further analyses since it captured all the key interactions between the chromatin marks, and because larger numbers of states did not capture sufficiently distinct interactions. The trained model was then used to compute the posterior probability of each state for each genomic bin in each reference epigenome. The regions were labelled using the state with the maximum posterior probability.

'Expanded' 18-state model. A second 'expanded' model applicable to 98 epigenomes that also have an H3K27ac ChIP-seq data set was learned by virtually concatenating consolidated data corresponding to the core set of five chromatin marks and H3K27ac. The model was trained on 40 high-quality epigenomes using the same parameters as those used for the primary model (Supplementary Table 1; Sheet QCSummary). We trained several models with the number of states ranging from 15 states to 25 states. An 18-state model was used for further analyses (Extended Data Fig. 2c) based on similar considerations.

State labels, interpretation and mnemonics. To assign biologically meaningful mnemonics to the states, we used the ChromHMM package to compute the overlap and neighbourhood enrichments of each state relative to various types of functional annotations (Fig. 4b, c, f and Extended Data Fig. 2b, c and Supplementary Fig. 2).

For any set of genomic coordinates representing a genomic feature and a given state, the fold enrichment of overlap is calculated as the ratio of 'the joint probability of a region belonging to the state and the feature' versus 'the product of independent marginal probability of observing the state in the genome' times 'the probability of observing the feature', namely the ratio between the (number of bases in state AND overlap feature)/(number of bases in genome) and the [(number of bases overlap feature)/(number of bases in genome) \times (number of bases in state)/(number of bases in genome)]. The neighbourhood enrichment is computed for genomic bins around a set of single-base-pair anchor locations such as transcription start sites.

For the overlap enrichment plots in the figures, the enrichments for each genomic feature (column) across all states is normalized by subtracting the minimum value from the column and then dividing by the max of the column. So the values always range from 0 (white) to 1 (dark blue); that is, it's a column-wise relative scale. For the neighbourhood positional enrichment plots, the normalization is done across all columns; that is, the minimum value over the entire matrix is subtracted from each value and divided by the maximum over the entire matrix.

The functional annotations used were as follows (all coordinates were relative to the hg19 version of the human genome): (1) CpG islands obtained from the UCSC table browser. (2) Exons, genes, introns, transcription start sites (TSSs) and transcription end sites (TESs), 2-kb windows around TSSs and 2-kb windows around TESs based on the GENCODEv10 annotation (<http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeGencodeV10/>) restricted to GENCODE biotypes annotating long transcripts. (3) Expressed and non-expressed genes, their TSSs and TESs. Genes were classified into the expressed or non-expressed class based on their RNA-seq expression levels in the H1-ES cells (Fig. 4c) and GM12878 (Extended Data Fig. 2b) cell lines. A gaussian mixture model with two components was fit on expression levels of all genes to obtain thresholds for the two classes. (4) Zinc finger genes (obtained by searching the ENSEMBL annotation for genes with gene names starting with ZNF). (5) Transcription factor binding sites (TFBS) based on ENCODE ChIP-seq data in the H1-ES cell line. The uniformly processed transcription factor ChIP-seq peak locations were downloaded from the ENCODE repository: <http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgTfbsUniform/>. We also computed percentage transcription factor binding site coverage for state calls in the GM12878 and K562 cell lines using corresponding transcription factor ChIP-seq data from ENCODE which matched and supported the mnemonics and state interpretations obtained from the H1 cell line (Supplementary Fig. 2). (6) Conserved GERP elements based on 34 way placental mammalian alignments <http://mendel.stanford.edu/SidowLab/downloads/gerp/> (Supplementary Fig. 3). (7) Enrichment for conserved GERP elements subtracting parts of the above-mentioned GERP elements that overlap exons.

Comparison to chromatin states learned on individual epigenomes. We also learned independent 15-state models individually on each of the 127 epigenomes using the core set of 5 marks and the same parameter settings as for the primary model. To compare the individual models to the joint 15-state primary model, we stacked the emission vectors for all states from all the models and hierarchically clustered them using Euclidean distance and Ward linkage (Extended Data Fig. 2a). The individual epigenome models consistently and repeatedly identified states that were also recovered by the joint model (Extended Data Fig. 2a). Two additional clusters which included states recovered by the independent models learned in individual cell types, but not recovered in the joint model, were HetWk, characterized

by weak presence of H3K9me3, and Rpts, characterized by presence of H3K9me3 along with a diversity of other marks, which was enriched in a large number of repeat elements.

Expanded chromatin states using large numbers of histone marks. For each of the seven deeply profiled reference epigenomes (Fig. 2j) we independently learned chromatin states on observed data for all available histone modifications or variants, and DNase in the reference epigenome. The same binarization and model learning procedure was followed as for the core set of 5 marks. We chose to consistently focus on a larger set of 50-states to capture the additional state distinctions afforded by using additional marks (Supplementary Fig. 4). Enrichments for annotations, including some of those described above for the 15-state model, were computed using ChromHMM. The HiC domains were obtained from ref. 107; the lamina-associated domains are described below; conserved element sets were the hg19 lift-over from ref. 73; repetitive element definitions were from RepeatMasker.

Relationship between histone marks, methylation and DNase. The distribution of DNA methylation (per cent CpG methylation from WGBS data) and DNA accessibility (DNase-seq $-\log_{10}(P \text{ value})$ signal confidence scores) was computed using regions belonging to each of the 15 chromatin states based on the core set of 5 marks and the 18 chromatin states from the expanded model across all reference epigenomes for which these data sets were available (Fig. 4d, e).

CpGs with a minimum read coverage of 5 were used to calculate the average methylation percentages within genomic regions labelled with each chromatin state from the 15-state primary model and 18 state expanded model. Only regions containing more than 3 CpGs with at most 200 bp between consecutive CpGs were used. Plots were generated using ggplot2 package for R (v.3.02). The average methylation levels for the chromatin states across DNA methylation platforms (WGBS, RRBS and mCRF) were analysed using Standard Least Square models in JMP (v.11.0; SAS Ins.). The model included the platforms (3 levels), chromatin states (15 levels) and the interactions (Extended Data Fig. 4).

Calling of lamina-associated domains. Genome-wide DamID binding data for human lamin B1 in SHEF-2 ES cells were obtained from GEO series GSE22428 (ref. 63). Lamina associated domains were determined using a similar method to the one described in ref. 64. First, hg18-based data coordinates were converted to hg19-based coordinates using UCSC's liftOver tool. Data were smoothed using a running median filter with a window size of 5 probes, after which domains were detected by estimating border and domain positions and comparing these to domains defined on 100 randomized instances of the same data set. Parameters are chosen such that the false discovery rate (FDR) for detected domains is 1%.

Chromatin state variability. For each state s for the core 15-state joint model we computed the number of genomic bins that were labelled with that state in at least one epigenome (G_s). From among these bins we counted the number of bins ($g_{s,i}$) that were labelled as being in state s in exactly i epigenomes ($i = 1 \dots 127$). We converted these counts to fractions ($g_{s,i}/G_s$) and computed the cumulative fraction that is consistently labelled with the same chromatin state in at most N epigenomes ($N = 1 \dots 127$). States whose cumulative fractions rise faster than others represent those that are less constitutive (more variable). We repeated the same procedure restricted to 43 high-quality and non-redundant Roadmap epigenomes (using only 1 representative epigenome from those corresponding to ES cells, iPS lines or epigenomes for the same tissue type from different individuals and excluding ENCODE cell lines) (Supplementary Table 1, Sheet VariationAnalysis) (Supplementary Fig. 6a). Analogous analysis was performed on states from the 18-state expanded model (Extended Data 5a and Supplementary Fig. 6b).

The observed cumulative fractions of cell-type specificity are a function of the composition of cell types in the compendium and do depend to some extent on the variability of data quality for the different marks. For example, the enhancer mark (H3K4me1) does have a much better signal-to-noise ratio than the transcribed mark (H3K36me3). One might expect this to result in more spurious variation of states associated with the transcribed mark. However, contrary to this expectation, the cumulative fractions for states involving only the transcription mark (Tx and TxWk) and not the enhancer mark indicate that these states are in fact less variable and more constitutive across cell types. On the other hand, all states composed of the enhancer mark (H3K4me1), irrespective of whether they do (TxFlnc, EnhG) or do not (EnhBiv, Enh, BivFlnc, TssAFlnk) include the transcription mark (H3K36me3), are far more cell-type specific. These observations indicate that the increased variability of states is largely due to the enhancer mark (H3K4me1) than the transcribed mark (H3K36me3). As replicates are not available in all epigenomes, we did not correct for inter-replicate variation in this analysis, but in the state-switching analysis below we utilize samples from the same tissue as quasi-replicates.

Chromatin state switching. To avoid spurious switching due to differences in data quality, we restricted this analysis to chromatin states from the 43 high-quality and non-redundant Roadmap epigenomes (see above). Using the 15 state primary model, we computed the empirical switching frequency of any pair of states across all pairs of 43 epigenomes. For a given pair of states A and B, we counted the number of

genomic bins that were labelled as (A,B) or (B,A) in all pairs of epigenomes. The switching frequency matrix (which is symmetric) was then row-normalized to convert the switching frequencies to switching probabilities. This is done to avoid a dependence on the total number of epigenomes. Also, the switching probabilities unlike switching frequencies are not dominated by states that are highly prevalent (for example, quiescent state). Supplementary Fig. 7b shows the empirical switching probabilities for all pairs of states across the 43 epigenomes. To differentiate between chromatin state dynamics across tissues (inter-tissue) relative to variation of states across individuals or replicates from the same tissue (intra-tissue), we also computed analogous switching frequencies by restricting to subgroups of epigenomes from the same tissue type (Supplementary Table 1, Sheet VariationAnalysis). The frequencies were added across all sub-groups and then row-normalized to switching probabilities. Supplementary Fig. 7a shows the intra-tissue switching probabilities. We then computed the relative enrichment of state switches as the \log_{10} ratio of inter-tissue switching probability across the 43 epigenomes relative to the intra-tissue switching probabilities (Fig. 5c). We repeated this analysis on the 16 ENCODE cell lines and obtained similar conclusions regarding relative enrichment of state switches (Supplementary Fig. 7c). Analogous analyses were performed using the 18-state expanded model in Roadmap Epigenomics samples (Extended Data Fig. 5c) and ENCODE samples (Supplementary Fig. 7d).

Large-scale chromatin structure. To study large-scale chromatin structure we first calculated ChromHMM (15-state model) state frequencies identified in 200-bp genome-wide bins across 127 epigenomes. Then we averaged state frequencies over the 2-Mb genomic regions, thus defining vectors of length 1,458 for each state. The unsupervised clustering of a $15 \times 1,458$ matrix (using Pearson correlation as a similarity measure and complete linkage) revealed 11 distinct genomic clusters enriched in different subsets of chromatin states (Fig. 5d, top heat map). Clusters had different sizes, with the smallest one (c1) containing only 27 bins, while the largest cluster (c9), occupied predominantly by a 'quiescent' state for all epigenomes, had 377 bins. For each 2-Mb bin in each cluster we calculated average gene density, lamin B1 signal (see section 4 above) and overlap with different cytogenetic bands (Fig. 5d, bottom, which displays also average levels across each cluster). We also show chromosomal locations of the clusters as well as distributions of CpG island frequency across the 2-Mb bins in each cluster (Extended Data Fig. 5d).

DMR calls across reference epigenomes. As a general resource for epigenomic comparisons across all epigenomes for which DNA methylation data is available, we defined DMRs using the method of Lister *et al.*¹⁰⁸, combining all differentially methylated sites (DMSs) within 250-bp of one another into a single DMR and excluded any DMR with less than 3 DMSs. For each DMR in each sample, we computed its average methylation level, weighted by the number of reads overlapping it¹⁰⁹. This resulted in a methylation level matrix with rows of DMRs and columns of samples.

DMRs in hESC differentiation (Fig. 4h). For analysing differentiation of hESCs in Fig. 4h, we used a second set of DMRs. We used a pairwise comparison strategy between ES cells and three *in vitro* derived cell types representative of the three germ layers (mesoderm, endoderm, ectoderm) and performed DMR calling as previously described⁵³. Only DMRs losing more than 30% methylation compared to the ES cell state at a significance level of $P \leq 0.01$ were retained. Subsequently, we computed weighted methylation levels for all three DMR sets across HUES64, mesoderm, endoderm and ectoderm as well as three consecutive stages of *in vitro* derived neural progenitors (please see accompanying paper⁵³ for details on the cell types). Finally, we plotted the corresponding distribution using the R function *vioplot* in the *vioplot* package. In order to identify potential regulators associated with the loss of DNA methylation at these regions, we determined binding sites of a compendium of transcription factors profiled in distinct cell lines and types that overlapped with each set of hypomethylated DMRs⁵¹. Next, we determined a potential enrichment over a random genomic background by randomly sampling 100 equally sized sets of genomic regions, respecting the chromosomal and size distribution of the different DMR sets and determined their overlap with the same transcription factor binding site compendium to estimate a null distribution. Only transcription factors that showed fewer binding sites across the control regions in 99 of the cases were considered for further analysis. Next, we computed the average enrichment over background for each transcription factor with respect to the 100 sets of random control regions for each germ layer DMR and report this enrichment level in Fig. 4h right, where we capped the relative enrichment at 12.

Additional DMR calls. For studying breast epithelia differentiation, DMRs were called from WGBS, requiring at least five aligned reads to call differentially methylated CpG, and at least three differentially methylated CpGs within a distance of 200 bp of each other⁴⁵. For studying tissue environment versus developmental origin, DMRs were called from MeDIP and MRE data using the M&M algorithm⁵⁶.

DNA methylation variation. For variation in methylation of each chromatin state across epigenomes (Fig. 4g and Extended Data Fig. 4f), we first excluded any contiguous chromatin state region containing less than three CpG sites. Then, the mean

of the methylation level for all contained CpG sites was calculated for each region, and for each epigenome density values were calculated for these mean methylation values between 0% and 100%, with density values estimated over $n = 1,000$ points with a gaussian kernel, with the default 'nrd0' bandwidth from the R stats package density function. Finally, for each chromatin state, we plotted the $\ln(\text{density} + 1)$ for each epigenome as rows, with the colour scale set with white as the minimum $\ln(\text{density} + 1)$ value and red, green, or blue, for WGBS, mCRF and RRBS, respectively, set as the maximum $\ln(\text{density} + 1)$ value in the matrix. Rows were ordered by the epigenomic lineage and grouping ordering shown in Fig. 2a. In Extended Data Fig. 4f, epigenomes were first grouped by methylation platform, and then ordered by Fig. 2a within each platform. The chromatin state methylation profiles in the cell lines versus primary cells/tissue cells were analysed using a mixed model with repeated measures. Overall effect of the group (cell lines versus primary cells/tissue cells) was tested using epigenomes within group as the error term. Testing for group effect was performed for each of the 15 chromatin states, resulting in a Bonferroni correction on the P values for the 15 tests.

Identifying coordinated changes in chromatin marks during development. To identify patterns of coordinated changes of histone marks over enhancers during heart muscle development, we compared ES cells, mesendoderm cells, and left ventricle tissue⁵⁷. We identified relevant enhancers as those that show changes in at least one histone mark between a specific cell type cluster (heart muscle in our case) and other cell types using LIMMA (Linear Model for Microarray Analysis). We applied FDR-corrected P value significance threshold of 0.05 to obtain cluster-specific enhancers. For each tissue type (heart muscle in our case) we then clustered the enhancers into five clusters (C1–C5) based on their multi-mark epigenomic profiles using the k-means algorithm implemented in the Spark tool (Fig. 4i). The tools used to generate Fig. 4i are integrated into the Epigenomic Toolset within the Genboree Workbench and are accessible for online use at <http://www.genboree.org>.

Clustering of epigenomes reveals common lineages and common properties. For each analysed mark, we calculated Pearson correlation values between all pairwise combinations of reference epigenomes using the mark's signal confidence scores ($-\log_{10}(\text{Poisson } P \text{ value})$) within 200 bp of the genomic regions deemed relevant for that mark. Relevance of regions is determined by whether a region was called in a particular (mark-matched) chromatin state with posterior probability of >0.95 in any of the reference epigenomes. For H3K4me1, H3K27ac and H3K9ac we used state Enh; for H3K4me3 state TssA; for H3K27me3 state ReprPC; for H3K36me3 state Tx; and for H3K9me3 state Het, unless otherwise noted (all based on the 15-state core model).

The resulting correlation matrices were used as the basis for a distance matrix for complete-linkage hierarchical clustering, followed by optimal leaf ordering¹¹⁰. Bootstrap support values are derived from 1,000 random samplings with replacement from all regions considered for a particular mark and a bootstrap tree was estimated for each resampling. The bootstrap support for a branch corresponds to the fraction of bootstrapped trees that support the bipartition induced by the branch.

In parallel to this, all correlation matrices mentioned above were used to perform Multi-Dimensional Scaling analyses using R.

Delineation of DNase I-accessible regulatory regions. For each of the 39 Roadmap reference epigenomes with DNase data, peak positions are combined across reference epigenomes by defining peak island areas, defined by stacking all DNase peak positions across epigenomes, and considering the full width at half maximum (FWHM). Note that for this we are only considering peak locations, not intensities. The goal of this is to obtain an estimate of the area of open chromatin, not to quantify the level of 'openness', as these data are not available for all reference epigenomes. In cases when peak islands overlap, they are merged because it means that the original DNase peak area populations overlap at least for half of the epigenomes with DNase peaks in that area (given the FWHM approach). Peak island summits are defined as the median peak summit of all peak island member DNase peaks. This results in a total of 3,516,964 DNase enriched regions across epigenomes.

We then annotate each of the $\sim 3.5\text{M}$ DNase peaks with the chromatin states they overlap with in each of the 111 Roadmap reference epigenomes, using the core 15-state chromatin state model, and focusing on states TssA, TssAFlnk and TssBiv for promoters, and EnhG, Enh and EnhBiv for enhancers, and state BivFlnk (flanking bivalent Enh/Tss) for ambiguous regions. Out of these, $\sim 2.5\text{M}$ regions are called as either enhancer or promoter across any of the 111 Roadmap reference epigenomes. Note that because DNase data are not available for all Roadmap epigenomes, the set of regulatory regions defined may exclude DNase regions active in cell types for which DNase was not profiled (Fig. 2g). Although most regions are undisputedly called exclusively promoter or enhancer, there are 535,487 regions that needed further study to decide whether they should be called promoters, enhancers, or both ('dyadic'). We arbitrate on these regions by first clustering them (using the methods in the following section) with an expected cluster size of 10,000

regions, and then for each cluster calculating (a) the mean posterior probabilities for promoter and enhancer calls separately, and (b) the mean number of reference epigenomes in which regions were called promoter or enhancer. Clusters of regions for which the differences in mean posterior probabilities (a) is smaller than 0.05, or for which the absolute \log_2 ratio of the number of epigenomes called as promoter or enhancer (b) is smaller than 0.05, are called true 'dyadic' regions, along with a small number of 'ambiguous' regions in state BivFlnk. Note that this particular clustering is only to arbitrate on these regions using group statistics instead of one-by-one; the final clusterings are described next. Overall, we define $\sim 2.3\text{M}$ putative enhancer regions (12.63% of genome), $\sim 80,000$ promoter regions (1.44% of genome) and $\sim 130,000$ dyadic regions (0.99% of genome), showing either promoter or enhancer signatures across epigenomes.

Clustering of DNase I-accessible regulatory regions to identify modules of co-ordinated activity. To cluster regulatory (that is, enhancer, promoter or dyadic) regions based on their activity patterns across all reference epigenomes, we expressed each region in terms of a binary vector of length $n \times s$, where n is the number of reference epigenomes (111) and s is the number of chromatin states considered. For enhancers and promoters, $s = 3$, as both of these types of regions are made up of 3 chromatin states in the 15-state ChromHMM model (enhancers, EnhG, Enh and EnhBiv; promoters, TssA, TssAFlnk and TssBiv).

The thus obtained binary matrices are subsequently clustered using a variation of a k-centroid clustering algorithm¹¹¹. Instead of Euclidean distance we use a Jaccard-index-based distance. This is done to be able to correctly cluster highly cell-type-restricted regions. From a computational point of view, we optimized the method to both deal with the size of the used data matrices and leverage their sparsity, to efficiently compute and update distances for matrices with sizes on the order of $10^6 \times 10^3$. The algorithm has been further modified to converge when less than 0.01% of cluster assignments change between iterations.

We selected the number of clusters k by tuning the expected number of regions within each cluster to be approximately 1,000 for promoter and dyadic regions, and approximately 10,000 for enhancer regions, given their much larger count (81,000, 129,000 and 2.3M for promoter, dyadic and enhancer, respectively). This results in a value of $k = 233$ for enhancer clusters (for $\sim 10\text{k}$ elements per cluster), and the algorithm converged on $k = 226$ non-empty clusters, which are used for subsequent analyses.

Clusters are visualized (Fig. 7a) by 'diagonalizing' when possible. First, 'ubiquitous' clusters (defined as having at least 50% of epigenomes with an enhancer/promoter density of $>25\%$) are shown. Then, the remaining clusters are ordered according to which epigenome has the maximum enhancer density.

Enrichment analyses of proximity to gene members of a catalogue of gene sets (Gene Ontology (GO), Human Phenotype Ontology (HPO)) have been performed using the GREAT tool⁵⁵. In particular, the GREAT web API was used to automatically submit region descriptions and retrieve results for subsequent parsing. We restricted ourselves to interpretation of results with an enrichment ratio of at least 2, and multiple hypothesis testing corrected P values <0.01 for both the binomial and the hypergeometric distribution based tests.

For visualization of a representative subset of enriched terms in Fig. 7b, c, we select representative terms for display (after diagonalizing the enrichment matrix by re-ordering the rows). We do this using a weighted bag-of-words approach to select highly enriched terms that contain many words that are over-represented in gene-set labels showing similar enrichment patterns. Briefly, sliding along the row names (gene-set terms) of the diagonalized enrichment matrices, we collect word counts and multiply these by integer-rounded $-\log_{10}(q \text{ values})$ obtained from GREAT. We do this in sliding windows of size 33 for Fig. 7b (resulting in 35 terms) and size 16 for Fig. 7c (resulting in 15 terms). For each word in a window, these values are expressed relative to the same words across all row names, registering to what extent they are over-represented. Each gene-set term in the window is then assigned a score based on the mean over-representation of all words it consists of. Lastly, gene sets are co-ranked based on this mean over-representation and their GREAT significance. The best-ranked gene set label is selected as the representative label for that window. All terms are shown in Supplementary Fig. 11d and are available for download at <http://compbio.mit.edu/roadmap>.

Predicting regulators active in each tissue, cell type and lineage. We collected 1,772 known transcription factor recognition motifs (position weight matrices) from primarily large-scale databases^{68,112–117} and measured their enrichment in the enhancers for each enhancer module compared to the union of the 226 enhancer modules (as described in refs 9,68) using a 0.3 conservation-based confidence cutoff^{70,73}. We clustered motifs using a 0.75 correlation cutoff resulting in 300 motif clusters⁶⁸ and selected for each motif cluster the motif with the highest enrichment in any enhancer module for further analysis.

We computed an expression score for each enhancer module and transcription factor as the Pearson correlation between the transcription factor expression across cell types with expression data (quantile-normalized $\log(\text{RPKM})$ with zeroes

replaced by $\log(0.0005)$ and the 'centre' of a module. For each enhancer module, its centre is defined as a vector of length 111, containing the fraction of regions in that module called as (any type of) enhancer in each of the 111 epigenomes analysed. This expression score is meant to act as the 'expression' of a transcription factor within a module of cell types. We then computed an expression-enrichment value for each transcription factor as the correlation of this expression score and the enrichment of the corresponding motif across enhancer modules. The top 40 motifs in terms of their absolute expression-enrichment correlation and the clusters with \log_2 enrichment or depletion of at least $\log_2 = 1.5$ for at least one motif are shown in Fig. 8 and Extended Data Fig. 8a (only one motif is shown in Fig. 8 for each factor).

We show all 84 motifs that were significantly enriched ($\log_2 \geq 1.5$) in any enhancer modules, across the full set of 226 enhancer modules (Supplementary Fig. 13a) and in the 101 modules in which they were significantly enriched (Extended Data Fig. 8a). Similarly, we show all 10 enriched motifs across the full set of 111 individual reference epigenomes (Supplementary Fig. 13b) and specifically in the 15 enriched epigenomes (Supplementary Fig. 13c). Lastly, we show all 19 enriched motifs across the full set of 17 tissue groups (Supplementary Fig. 13d), and specifically within the 10 groups that showed significant enrichments (Supplementary Fig. 13e).

For visualization of regulator–cell type links (Fig. 8), we computed edge weights between each cell type and motif using these motif-module enrichments. For each motif and cell type, we computed the sum across all modules of the product of the \log_2 motif enrichment and the value of the cell type within the module centre (only consider the highly associated cell types by replacing values < 0.7 with 0). We show all resulting edge weights of at least 1.5 and visualize the network using Cytoscape¹¹⁸.

Based on the same motif enrichment method mentioned above, we computed the motif enrichment in the tissue-specific Digital Genomic Footprinting (DGF) regions in each library. The tissue-specific DGF regions were identified by selecting the DGF region occurring in no more than 20 DGF libraries among 42 DGF libraries. To generate Extended Data Fig. 9b, we standardized the motif enrichment in each library into z-scores for each motif (row) and colour each DGF library (column) based on their tissue type.

DNA motif positional bias in digital genomic footprinting sites. We computed the positional enrichment of each driver motif (Extended Data 9c and 10) related to the digital genomic footprinting (DGF) sites in each cell type (Supplementary Table 5b). For each driver transcription factor motif, we generated two views corresponding to the motif position (the centre of the motif instance) relative to the centre of closest DGF site (centre view) and the motif position relative to the boundary of closest DGF site (boundary view). We only considered the motif instances with closest DGF site within 100 bp. For the centre view, we plotted the motif occurrence density versus the distance to the DGF centre for different cell types. For the boundary view, we considered the shortest distance between the centre of a motif instance and either side of DGF boundary, and gave a negative distance value for motif instances inside the DGF, and a positive distance value otherwise. Similar to the centre view, we plotted the motif density versus the derived distance value in the boundary view for each cell type.

To access the significance of the motif concentration within DGF in each cell type, we computed the DGF enrichment ratio as the ratio between the number of motif instances with distance less than 20 bp to the DGF centre and that number in the immediate flanking window, that is, the number of motif instances with distance to the DGF centre larger than 20 bp and smaller than 40 bp. As control, we randomly sampled the same number of motif instances from the shuffled versions of the given motif, and obtained the DGF enrichment ratio for the shuffled motif instances. The DGF enrichment ratio of the true motif is further converted to z-score by mean and standard deviation from the DGF enrichment ratios of shuffled motif from 1,000 times random sampling. Then the adjusted *P* value is further computed from z-score and Bonferroni correction for number of cell types.

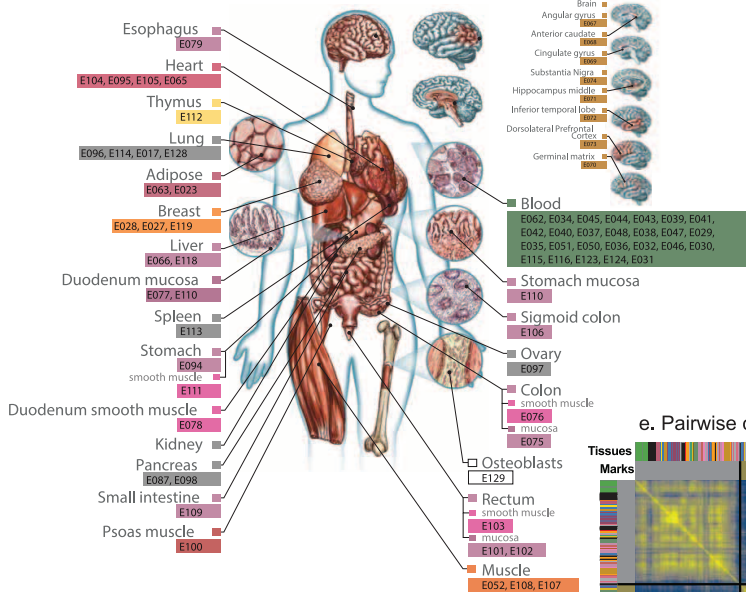
Comparing DGF with DNA motifs that are predictive of epigenomic modification. The motifs that were predictive of epigenomic modifications⁷¹ were compared to DGF in Supplementary Table 5a. This was done in three cell types where both DGF and predictive motifs were available: 'H1 BMP4 derived mesendoderm cultured cells' (E004), 'H1 BMP4 derived trophoblast cultured cells' (E005), and 'H1 derived mesenchymal stem cells' (E006). The motifs that were predictive of

the following seven inputs were considered: H3K27me3, H3K27ac, H3K9me3, H3K36me3, H3K4me1, H3K4me3 and DNA methylation valleys (DMV)¹¹. To identify overlaps the predictive motifs were scanned against the modification peaks of the corresponding modification and the location of the best match between motif and sequence was recorded. Then we counted the number of times the locations of the best motif matches overlapped a DGF by at least 1 bp. These counts were compared to the number of overlaps identified randomly, which was calculated by comparing DGF to random locations within the modifications peaks. The reported random frequency was the average of 100 repeats. To calculate the fold enrichment we divided the observed frequency by the random frequency.

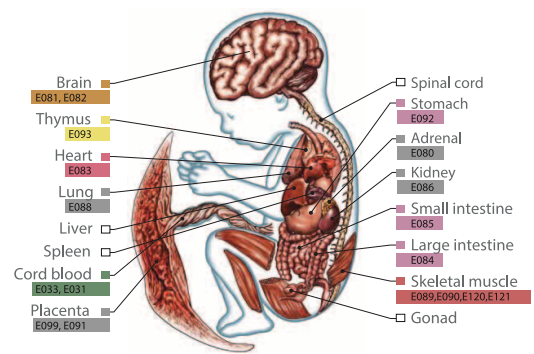
Tissue-specific activity of disease-associated regions. We tested the enrichment of SNPs from individual genome-wide association studies (GWAS) for the gapped peak call sets for histone marks H3K4me1, H3K4me3, H3K36me3, H3K9me3, H3K27me3, H3K9ac and H3K27ac as well as the DNase peak call set based on MACS2 in each reference epigenome where available. The SNPs used were curated into the NHGRI GWAS catalogue⁷⁵ and obtained through the UCSC Table Browser¹¹⁹ on 12 September 2014. We restricted the enrichment analysis to chromosomes 1–22 and chromosome X. We defined a study to be a unique combination of annotated trait and PubMed ID. To reduce dependencies between pairs of SNPs assigned to the same study, we pruned SNPs such that no two SNPs were within 1 Mb of each other on the same chromosome. The pruning procedure considered each SNP in ranked order of *P* value with the most significant coming first, and we retained a SNP if there was no already retained SNP on the same chromosome within 1 Mb. We computed hypergeometric *P* values for the enrichment of each pruned set of SNPs overlapping peak calls against the pruned GWAS catalogue as the background. We estimated separately for each mark a mapping from a *P* value to a false discovery rate across tests for all study and reference epigenome combinations by generating 100 randomized versions of the pruned GWAS catalogues, shuffling which SNPs were assigned to which study and computing the average fraction of reference epigenome–study combinations that reached that level of significance (in a continuous mapping of *P* values to FDR) using randomized catalogues divided by the number based on the actual GWAS catalogue.

104. John, S. *et al.* Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nature Genet.* **43**, 264–268 (2011).
105. Ernst, J. & Kellis, M. Interplay between chromatin state, regulator binding, and regulatory motifs in six human cell types. *Genome Res.* **23**, 1142–1154 (2013).
106. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods* **9**, 215–216 (2012).
107. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
108. Lister, R. *et al.* Global epigenomic reconfiguration during mammalian brain development. *Science* **341**, 1237905 (2013).
109. Schultz, M. D., Schmitz, R. J. & Ecker, J. R. 'Leveling' the playing field for analyses of single-base resolution DNA methylomes. *Trends Genet.* **28**, 583–585 (2012).
110. Bar-Joseph, Z., Gifford, D. K. & Jaakkola, T. S. Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics* **17** (suppl. 1), S22–S29 (2001).
111. Leisch, F. A toolbox for KK-centroids cluster analysis. *Comput. Stat. Data Anal.* **51**, 526–544 (2006).
112. Matys, V. *et al.* TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.* **31**, 374–378 (2003).
113. Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W. W. & Lenhard, B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* **32**, D91–D94 (2004).
114. Berger, M. F. *et al.* Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nature Biotechnol.* **24**, 1429–1435 (2006).
115. Berger, M. F. *et al.* Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell* **133**, 1266–1276 (2008).
116. Jolma, A. *et al.* DNA-binding specificities of human transcription factors. *Cell* **152**, 327–339 (2013).
117. Badis, G. *et al.* Diversity and complexity in DNA recognition by transcription factors. *Science* **324**, 1720–1723 (2009).
118. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
119. Karolchik, D. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32**, D493–D496 (2004).

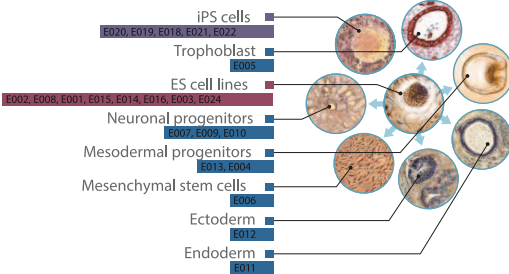
a. Primary tissues and cells - adult samples



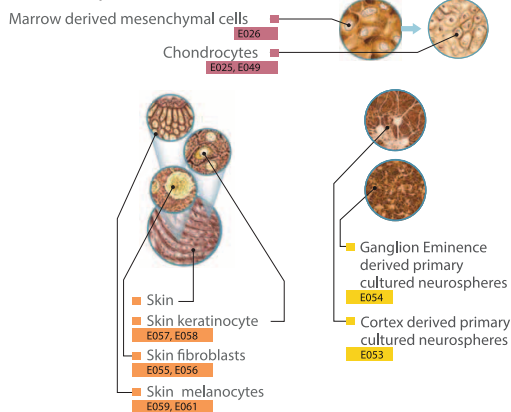
b. Primary tissues and cells - fetal samples



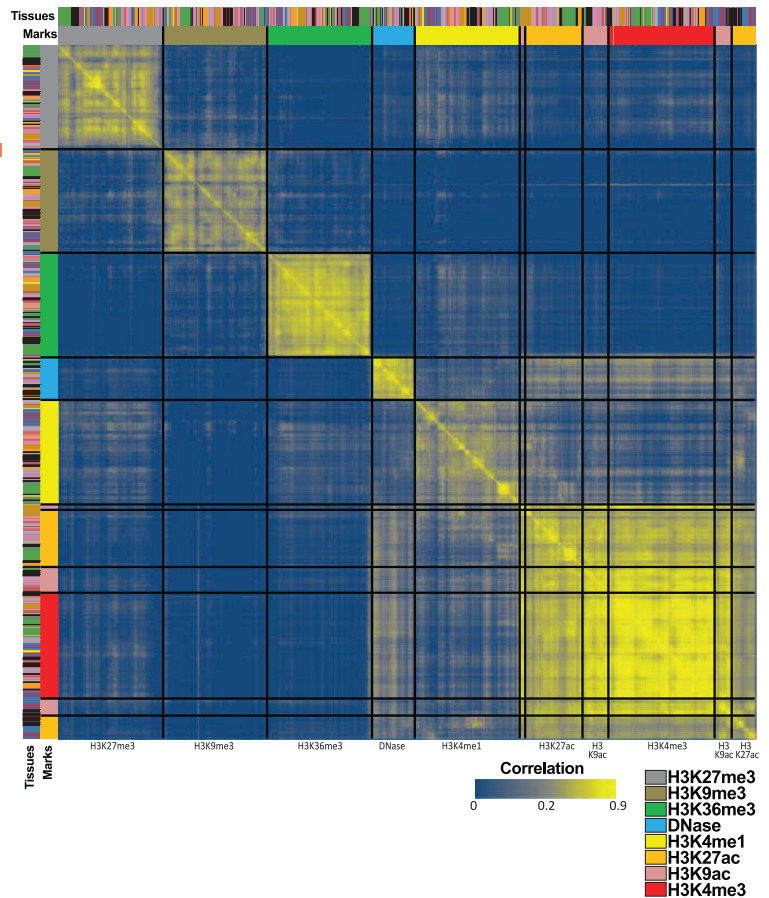
c. ES cells, iPSC, and ES cell-derived cells



d. Primary cultures

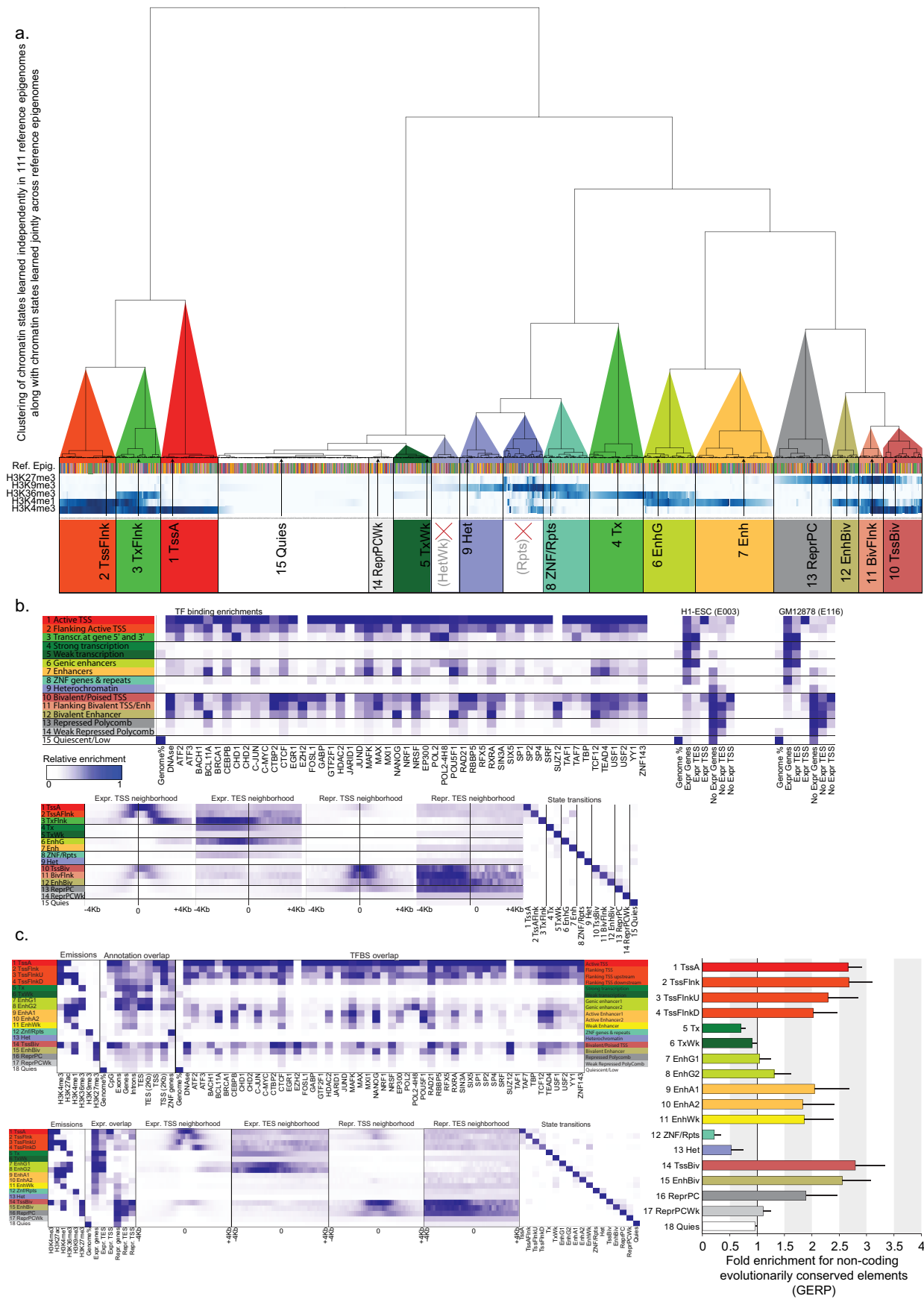


e. Pairwise correlations of all histone marks and DNA accessibility datasets



Extended Data Figure 1 | Tissues and cell types of reference epigenomes. Comprehensive listing of all 111 reference epigenomes generated by the consortium, along with epigenome identifiers (EIDs), including: (a) adult samples; (b) fetal samples; (c) ES cell, iPS cell and ES-cell-derived cells; and (d) primary cultures. Colours indicate the groupings of tissues and cell types (as in Fig. 2b, and throughout the manuscript). For five samples (adult osteoblasts, and fetal liver, spleen, gonad and spinal cord), no colour is present, indicating that these are not part of the 111 reference epigenomes (ENCODE 2012 samples, or not all five marks in the core set were present), but data sets from these samples are high quality and were sometimes used in companion paper analyses, and are publicly available. e. Assay correlations. Heat map of the

pairwise experiment correlations for the core set of five histone modification marks (H3K4me1, H3K4me3, H3K36me3, H3K27me3, H3K9me3) across all 127 reference epigenomes, the two common acetylation marks (H3K27ac and H3K9ac), and DNA accessibility (DNase) across the reference epigenomes where they are available. Yellow indicates relatively higher correlation and blue lower correlation. Rows and columns were ordered computationally to maximize similarity of neighbouring rows and columns (see Methods). All experiments for H3K9me3, H3K27me3, H3K36me3, DNase and H3K4me1 are consistently ordered into distinct and contiguous groups. For H3K4me3, H3K9ac and H3K27ac, experiments group primarily based on the mark, but in some cases, the correlations and ordering appear more cell-type driven.



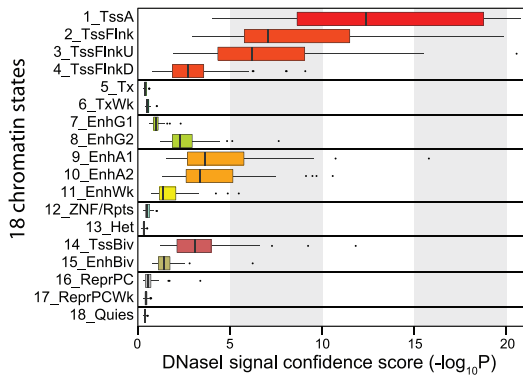
Extended Data Figure 2 | Chromatin state model robustness and enrichments.

a, Chromatin state model robustness. Clustering of 15-state 'core' chromatin state model learned jointly across reference epigenomes (Fig. 4a) with chromatin state models learned independently in 111 reference epigenomes. We applied ChromHMM to learn a 15-state ChromHMM model using the five core marks in each of the 111 reference epigenomes generated by the Roadmap Epigenomics program, and clustered the resulting 1,680-state emission probability vectors (leaves of the tree) with the 15 states from the joint model (indicated by arrows). We found that the vast majority of states learned across cell types clustered into 15 clusters, corresponding to the joint model states, validating the robustness of chromatin states across cell types. This analysis revealed two new clusters (red crosses) which are not represented in the 15 states of the jointly learned model: 'HetWk', a cluster showing weak enrichment for H3K9me3; and 'Rpts', a cluster showing H3K9me3 along with a diversity of other marks, and enriched in specific types of repetitive elements (satellite repeats) in each cell type, which may be due to mapping artefacts. This joint clustering also revealed subtle variations in the relative frequency of presence of H3K4me1 in states TxFlnk, Enh and TssBiv, and H3K27me3 in state TssBiv. Overall, this analysis confirms that the 15-state chromatin state model based on the core set of five marks provides a robust framework for interpreting epigenomic complexity across tissues and cell types.

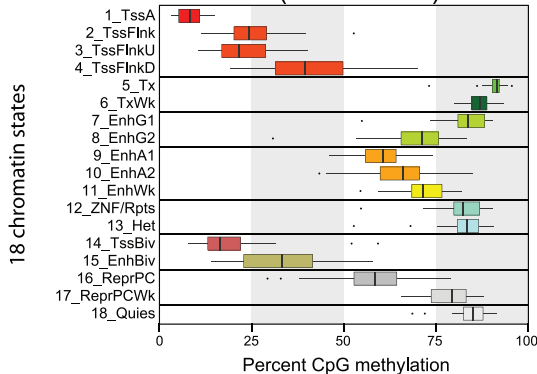
b, Enrichments for 15-state model based on five histone modification marks. Top left: transcription factor binding site overlap enrichments of 15 states in H1-ES cells from the 'core' model for transcription factor binding sites

(TFBS) based on ChIP-seq data in H1-ES cells. Transcription factor binding coverage for other cell types based on matched transcription factor ChIP-seq data are shown in Supplementary Fig. 2. Top right: enrichments for expressed and non-expressed genes in H1-ES cells and GM12878. Bottom: positional enrichments at the transcription start site (TSS) and transcription end site (TES) of expressed (expr.) and repressed (repr.) genes in H1-ES cells. Transition probabilities show frequency of co-occurrence of each pair of chromatin states in neighbouring 200-bp bins. **c**, Definition and enrichments for 18-state 'expanded' model that also includes H3K27ac associated with active enhancer and active promoter regions, but which was only available for 98 of the 127 reference epigenomes. Inclusion of H3K27ac distinguishes active enhancers and active promoters. Top: TFBS enrichments in H1-ES cells (E003) chromatin states using ENCODE transcription factor ChIP-seq data in H1-ES cells. Bottom: positional enrichments in H1-ES cells for genomic annotations, expressed and repressed genes, TSS and TES, and state transitions as in Extended Data Fig. 2b and Fig. 4a–c. Right: average fold-enrichment (colours bars) and standard deviation (black line) across 98 reference epigenomes (Supplementary Fig. 3d) for the fold enrichment for non-exonic genomic segments (GERP) in each chromatin state (rows) in the 18-state model. Excluding protein-coding exons (see Supplementary Fig. 3b versus Supplementary Fig. 3d), the TSS-proximal states show the highest levels of conservation, followed by EnhBiv and the three non-transcribed enhancer states. In contrast, Tx and TxWk elements are weakly depleted for conserved regions, and Znf/Rpts and Het are strongly depleted for conserved elements.

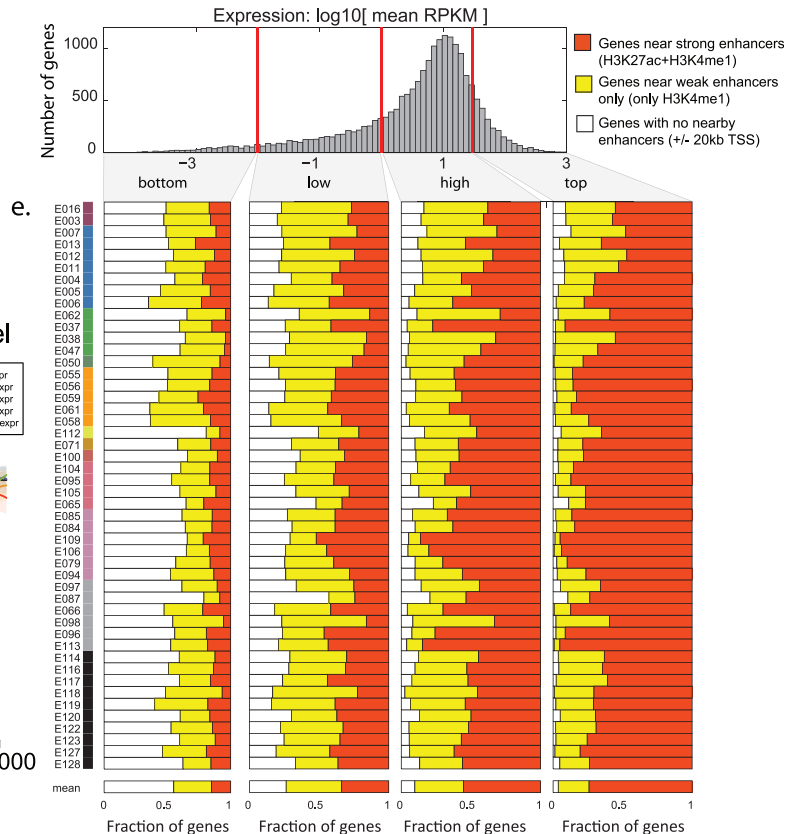
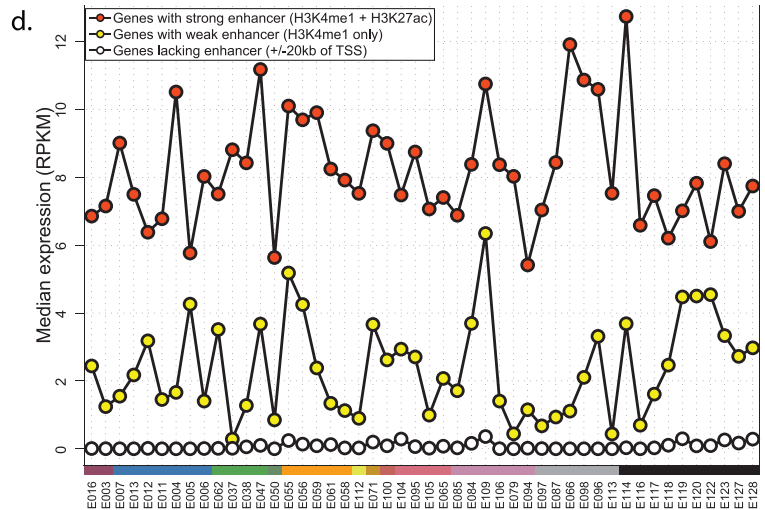
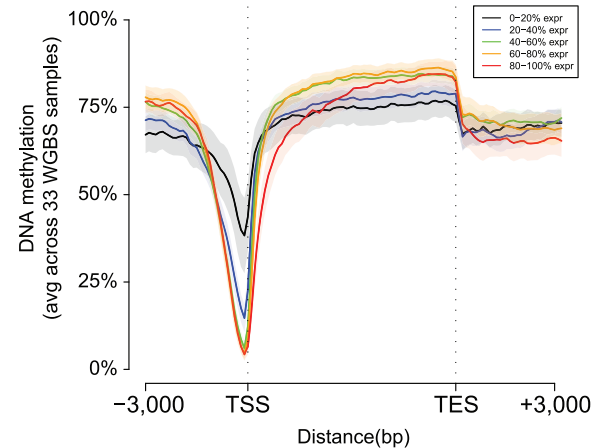
a. DNA accessibility (44 datasets)



b. WGBS enrichment (33 datasets)

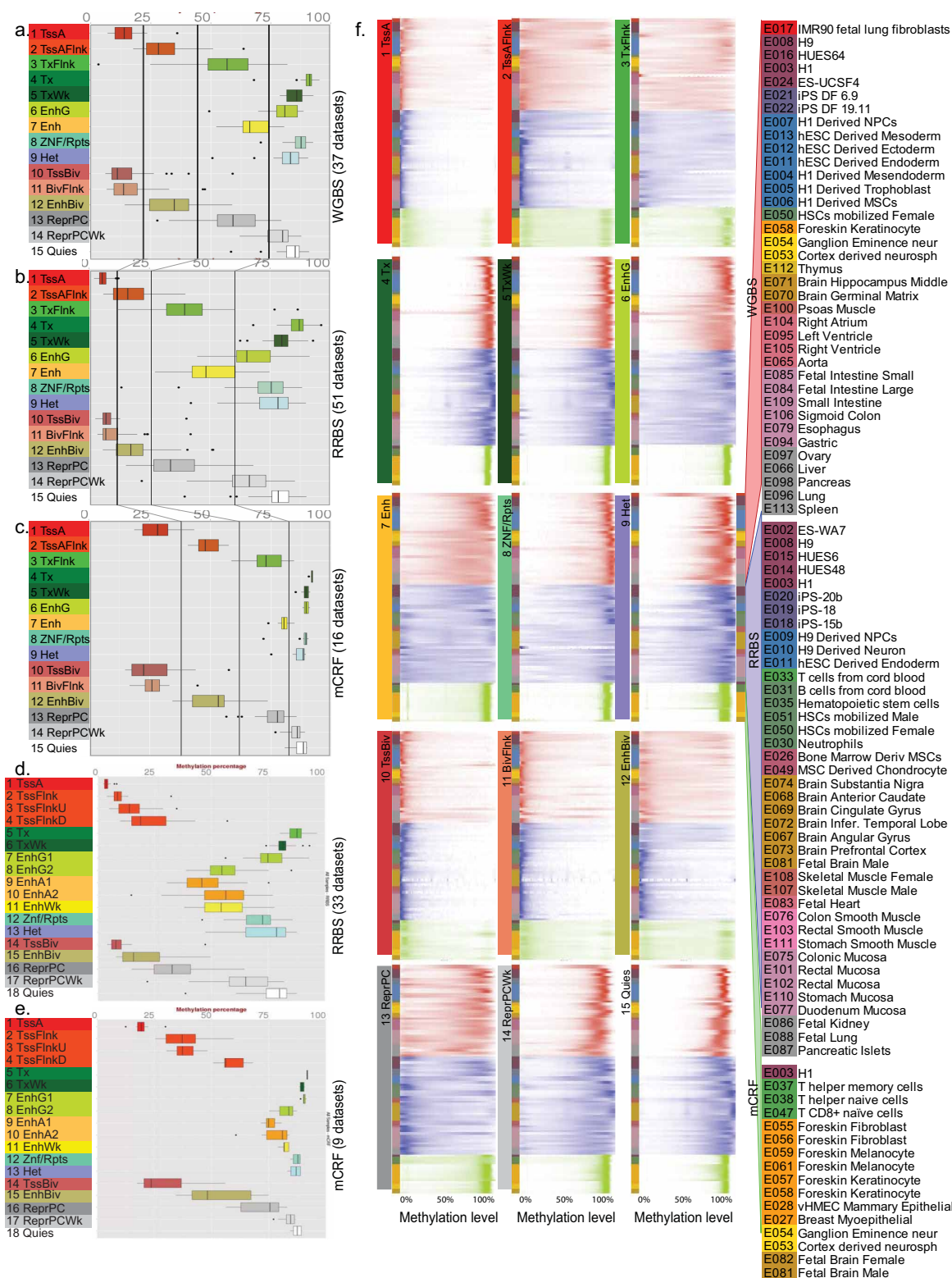


c. DNA methylation vs. gene expression level



Extended Data Figure 3 | Relationship between histone marks, DNA methylation, DNA accessibility and gene expression. **a**, H3K27ac-marked 'active' enhancers show higher levels of DNA accessibility, based on enrichment of DNase-seq signal confidence scores ($-\log_{10}(\text{Poisson } P \text{ value})$) for elements in each chromatin state in our extended 18-state model that includes the core five histone modification marks and H3K27ac, similar to Fig. 4e. **b**, Level of whole-genome bisulfite methylation for all chromatin states in the 18-state model shows that H3K27ac-marked 'active' enhancers associated with H3K27ac in addition to H3K4me1 show lower methylation levels, consistent with higher regulatory activity. The whiskers in **a** and **b** show $1.5\times$ interquartile range and the filled circles are individual outliers. **c**, DNA methylation levels for genes showing different expression levels. The depletion of DNA methylation in promoter regions, and the enrichment of DNA methylation in transcribed regions, are both more pronounced for highly expressed genes. The enrichment for high DNA methylation is more pronounced in the 3' ends

of the most highly expressed genes. **d**, Genes associated with active enhancer states have consistently significantly higher expression. 'Active enhancer' associated genes have at least one EnhA1 and/or EnhA2 ± 20 kb from TSS (18-state model). 'Weak-enhancer' genes are associated with EnhG1, EnhG2, EnhWk, EnhBiv. Lowest expression have genes that are not associated with any enhancer. Plots with red markers show median expression of genes associated with 'active' enhancers, yellow markers 'weak' enhancers, and white markers no association with any enhancer state. **e**, Higher-expression genes show greater association with H3K27ac-marked 'active' enhancers. Highly expressed genes are consistently more frequently associated with H3K27ac-marked active enhancers (EnhA1 and EnhA2) across all cell types. Fraction of genes associated with H3K27ac-marked 'active' enhancers (red), H3K27ac-lacking 'weak' enhancers only (yellow), or no enhancers (white) for genes of varying expression levels in each cell type with RNA-seq data.

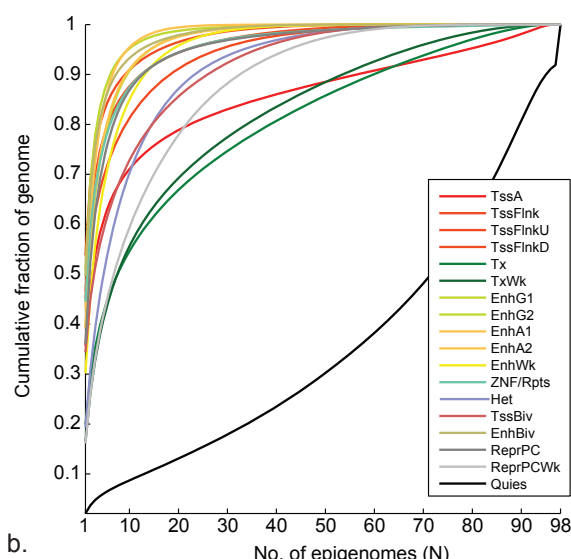


Extended Data Figure 4 | Methylation relationship with chromatin state.

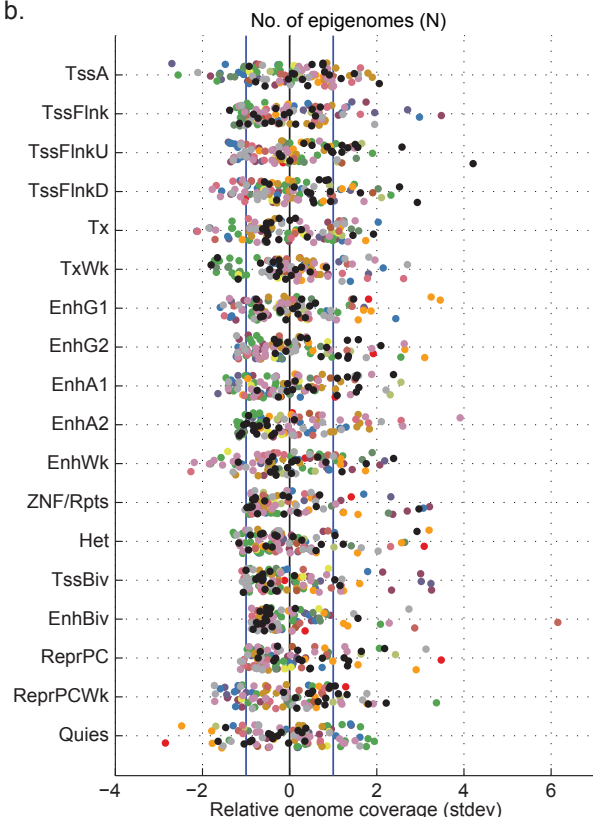
a–c, DNA methylation levels in 15-state model across technologies. We observed significant differences in the average methylation levels observed that were correlated with the different DNA methylation platforms used, but their relative relationships in average chromatin state methylation were conserved. Relative to WGBS (panel **a**, repeated from Fig. 4d for comparison purposes), RRBS (panel **b**) showed the lowest overall methylation levels (as expected given its CpG island enrichment), while mCRF showed the highest (panel **c**). This highlights the importance of recognizing and potentially correcting for DNA-methylation-platform-specific biases before performing integrative analysis. **d, e**, Distribution of DNA methylation levels measured using RRBS and mCRF

in 18-state model (defined in Extended Data Fig. 2c). WGBS is shown in Extended Data Fig. 3b. The whiskers in **a–e** show 1.5 \times interquartile range and the filled circles are individual outliers. **f**, DNA methylation variation across cell types. Density plots denote distribution of DNA methylation levels from 0% to 100% for each chromatin state across the 95 reference epigenomes profiled for whole-genome bisulfite (WGBS, red), reduced representation bisulfite (RRBS, blue), or MeDIP/MRE (mCRF, green). The respective colour (red, blue, or green) was set to the maximum $\ln(\text{density} + 1)$ value for each chromatin state and respective platform, with intermediate values coloured on a natural log scale. For each panel, the subset of reference epigenomes profiled using each technology are listed, using the colours, order, and abbreviations from Fig. 2.

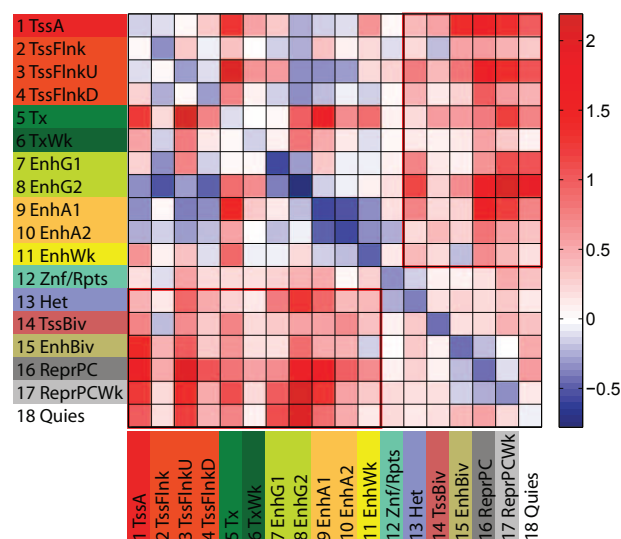
a. Chromatin state variability for 18-state expanded model across 98 epigenomes



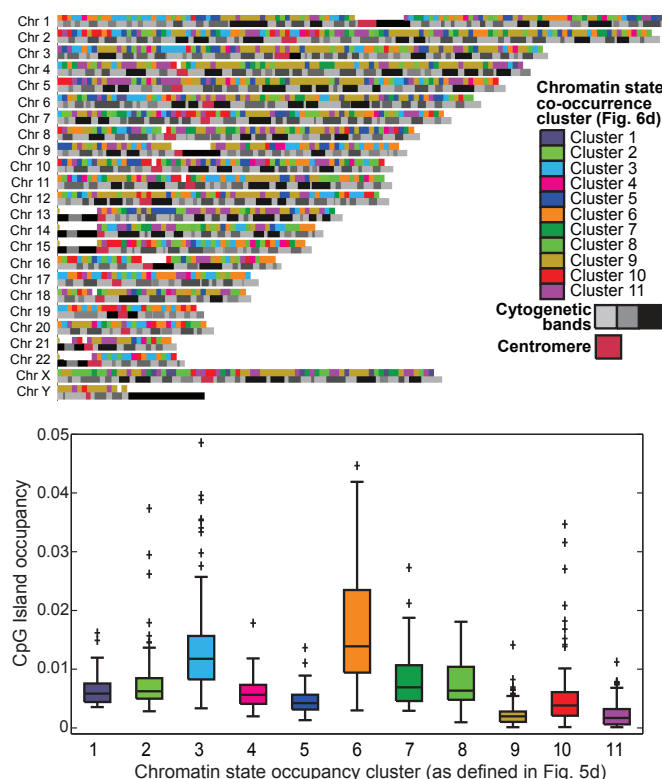
b.



c.

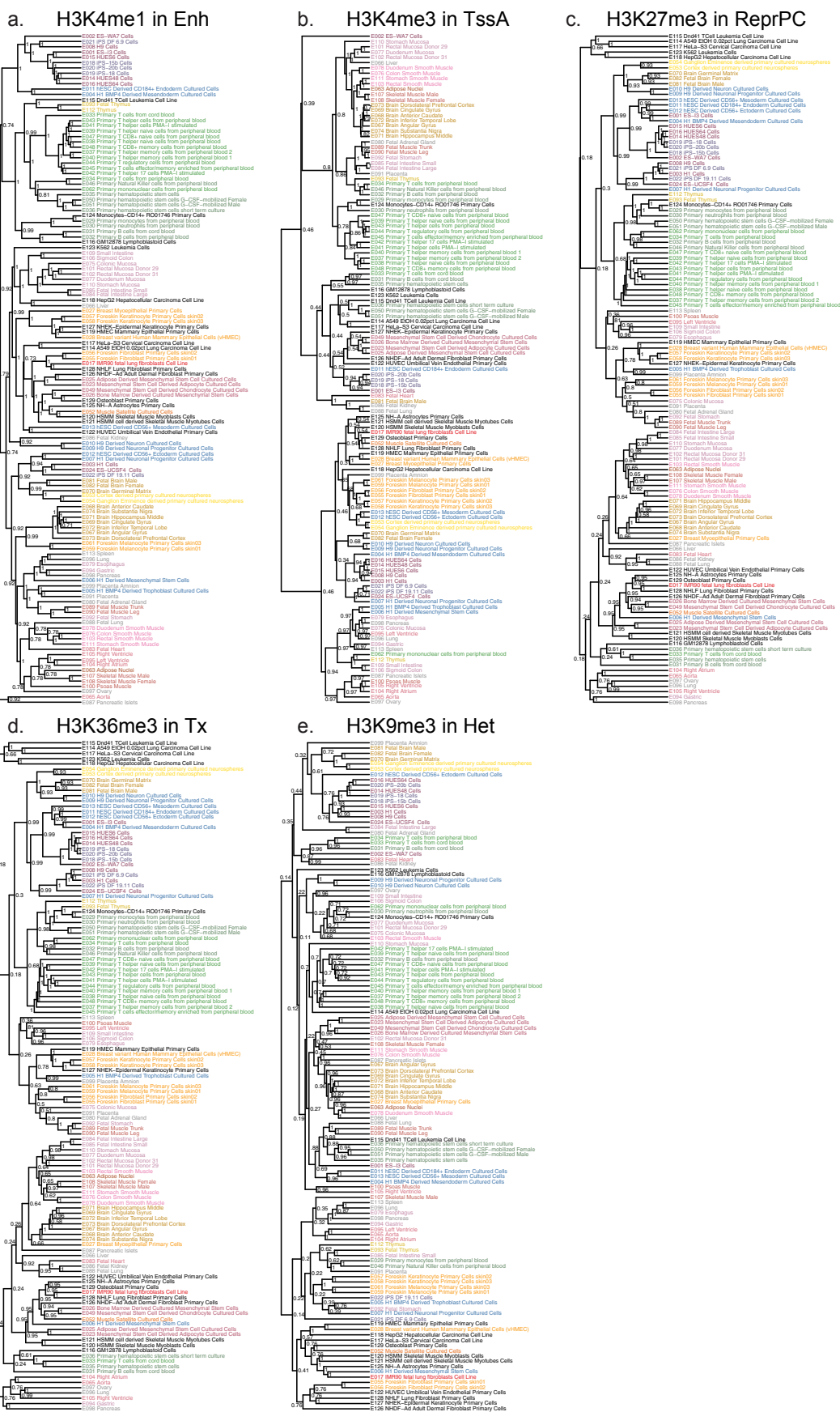


d.



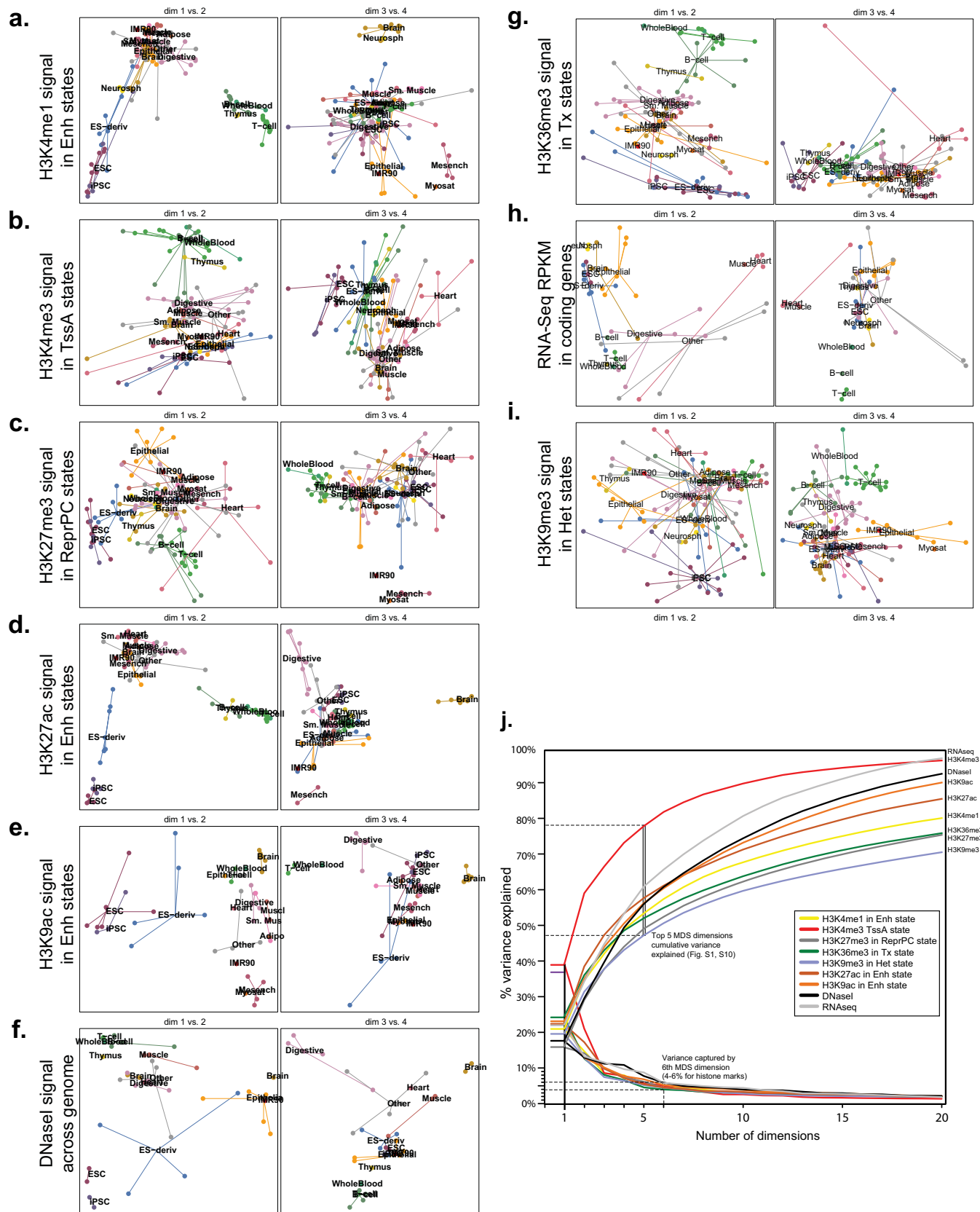
Extended Data Figure 5 | Chromatin state variability, switching and genomic coverage. **a.** Variability level for 18-state model. Chromatin state variability (similar to Fig. 5a), quantified based on the fraction of the genomic coverage (y axis) of each state (colour) that is consistently labelled with that state in at most N (ranging from 1 to 98) reference epigenomes, using the 18-state model learned based on 6 chromatin marks, including H3K27ac. **b.** Chromatin state over- and under-representation for 18-state expanded model. **c.** Log-ratio (\log_{10}) of chromatin state switching probabilities for the 18-state expanded model across 34 high-quality, non-redundant epigenomes that have H3K27ac data, relative to intra-tissue switching probabilities across replicates or samples from multiple individuals. **d.** Chromatin state coverage

grouped by epigenomic domains. Top: chromosome 'painting' of 11 clusters shown in Fig. 5d and discovered based on chromatin state co-occurrence at the 2-Mb scale across reference epigenomes. Bottom: enrichment of CpG islands in each cluster clearly showing higher CpG density 'active' clusters 3 and 6 comparing to passive clusters 9–11. Each box plot shows a distribution of CpG total occupancy in 2-Mb bins in each cluster (with box boundaries indicating 25th and 75th percentiles, the whiskers extend to the most extreme data points considered to not be outliers). Points are drawn as outliers if they are larger than $Q3 + 1.5 \times (Q3 - Q1)$ or smaller than $Q1 - 1.5 \times (Q3 - Q1)$, where $Q1$ and $Q3$ are the 25th and 75th percentiles, respectively.



Extended Data Figure 6 | Hierarchical clustering of epigenomes using diverse marks. a–e, Clustering of all 127 reference epigenomes, including ENCODE samples, using H3K4me1, H3K4me3, H3K27me3, H3K36me3 and H3K9me3 signal in Enh, TssA, ReprPC, Tx and Het chromatin states,

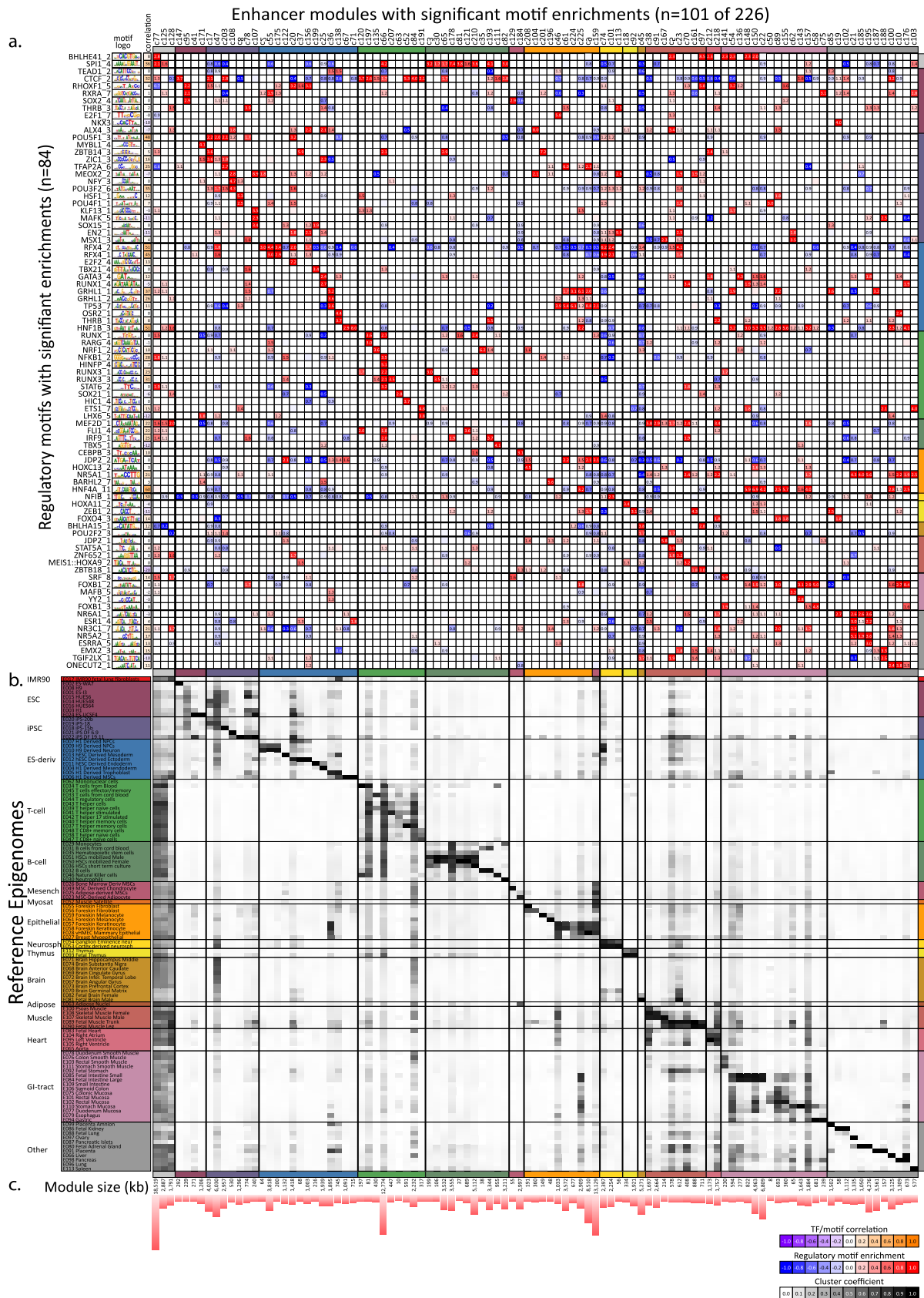
respectively. All panels show hierarchical clustering with optimal leaf ordering. Colours indicate sample groups, as defined in Fig. 2. Numbers on internal nodes represent bootstrap support scores over 1,000 bootstrap samples.



Extended Data Figure 7 | Multi-dimensional scaling (MDS) analysis.

a–i. MDS plots showing reference epigenome distances using similarity of different epigenomic marks in corresponding chromatin states. Reference epigenomes (dots) are coloured according to their group colouring defined in Fig. 2b. Thin lines connect same-group reference epigenomes. The first four axes of variation are shown in pairs. Marks are assessed in regions with relevant chromatin states (see

Methods). **j.** Variance explained by each MDS dimension. The first five dimensions shown in Supplementary Fig. 10 (Fig. 6b, c) explain between 45% and 80% of the total epigenome-to-epigenome variance for all histone modification mark correlations, and additional dimensions explain less than 10%. Only a few components of H3K4me3 in TssA chromatin states explains a much larger fraction of the variance than other marks, possibly due to its stability across cell types.

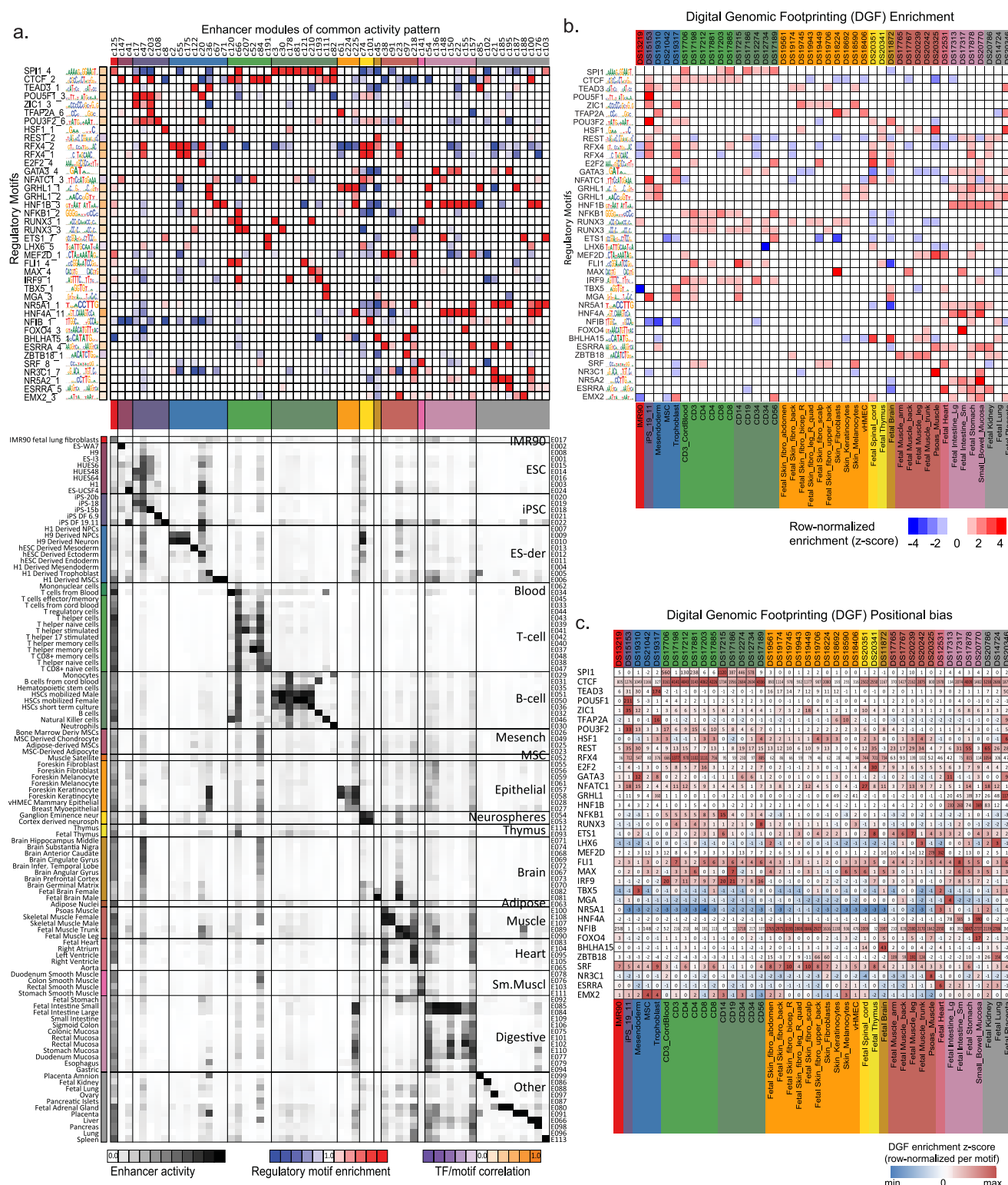


Extended Data Figure 8 | Regulatory motif analysis for modules.

Regulatory motifs enriched in enhancer modules. Enrichment (red) or depletion (blue) of regulatory motifs (rows) in the enhancer modules (columns) relative to shuffled control motifs. For each motif is shown the motif name, consensus logo, and correlation between regulator expression and module activity: positive correlation (orange) is indicative of activators, and negative

correlation (purple) indicates a repressive role for the factor. Only clusters with log enrichment or depletion of at least 1.5-fold for one motif are shown.

b. Average activity level of enhancers of each module in each reference epigenome (black, high; white, low). **c.** Total size of each enhancer module showing enrichment (in kb).



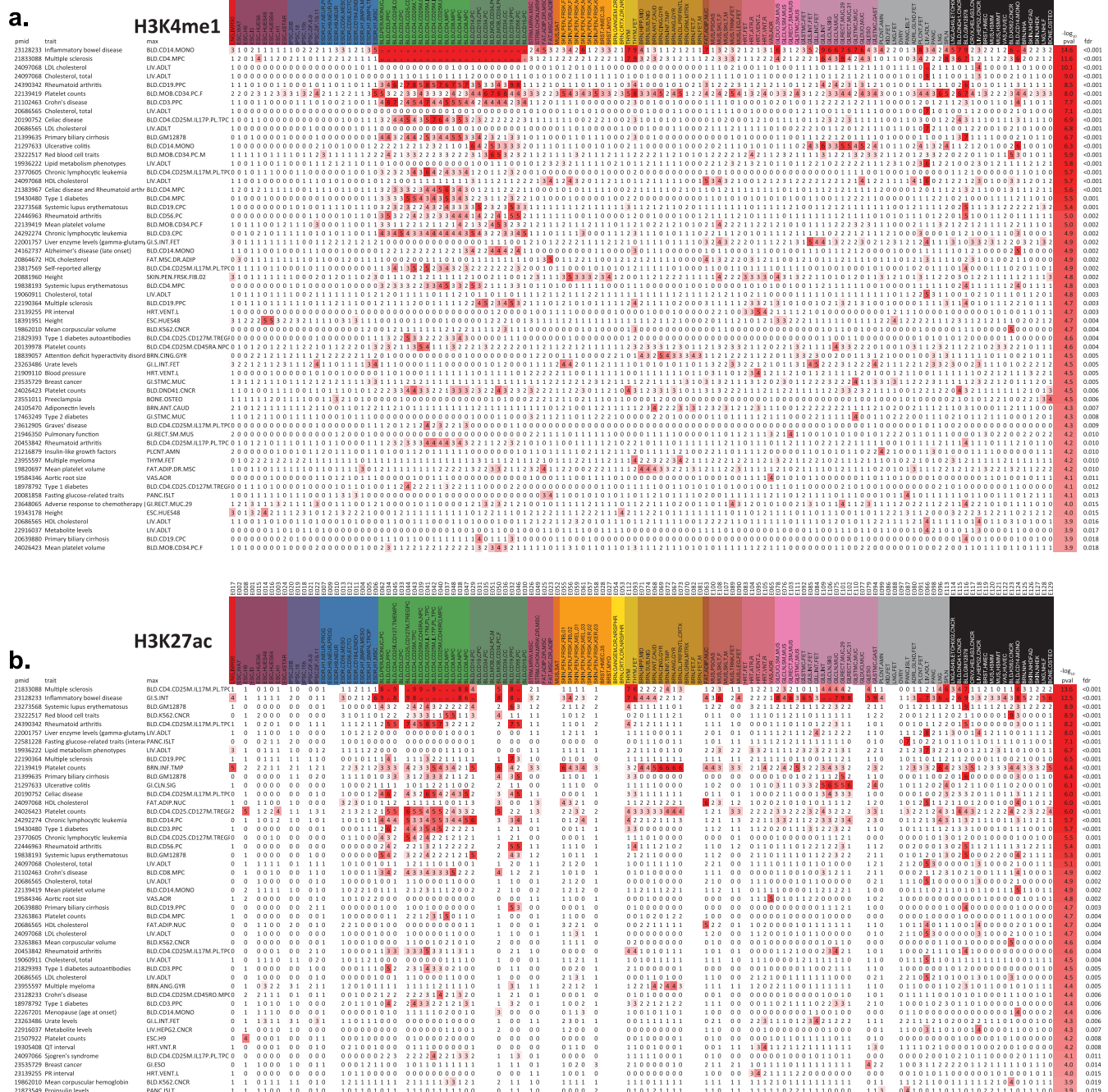
Extended Data Figure 9 | Regulatory motif enrichment, DGF enrichment and positional bias for predicted driver motifs. **a**, Regulatory motif enrichments for the 40 regulators showing the strongest absolute correlation between transcription factor expression and module activity. Of these, 36 were also recovered solely based on their motif enrichment scores (Extended Data Fig. 8), but 6 motifs showing significant and biologically relevant correlations were not discovered solely based on their motif enrichment (Esrra_4, Max_4, Mga_3, Nfatc1_3, Rest_2 and Tead3_1), illustrating the importance of studying motif enrichments in the context of transcription factor expression and enhancer activity patterns. **b**, Predicted driver regulatory motifs

are enriched in high-resolution DNase footprints. Enrichment of predicted driver motif instances (Fig. 8 and Extended Data Fig. 9a) in 42 high-resolution (6–40 bp) DGF libraries from deeply sequenced DNase data sets⁵⁹ shows consistent tissue preferences in matching cell types. For example, POU5F1 in iPS cells, HNF1B and HNF4A1 in digestive tissues, RFX4 in neural lineages, MFE2B in muscle. **c**, Matrix of significant positional bias across factors and cell types. For each DGF data set (columns), positional bias score (heat map) of predicted driver regulatory motifs (rows) found to be significantly enriched (Fig. 8 and Extended Data Fig. 9a) in enhancer modules (Fig. 7a).



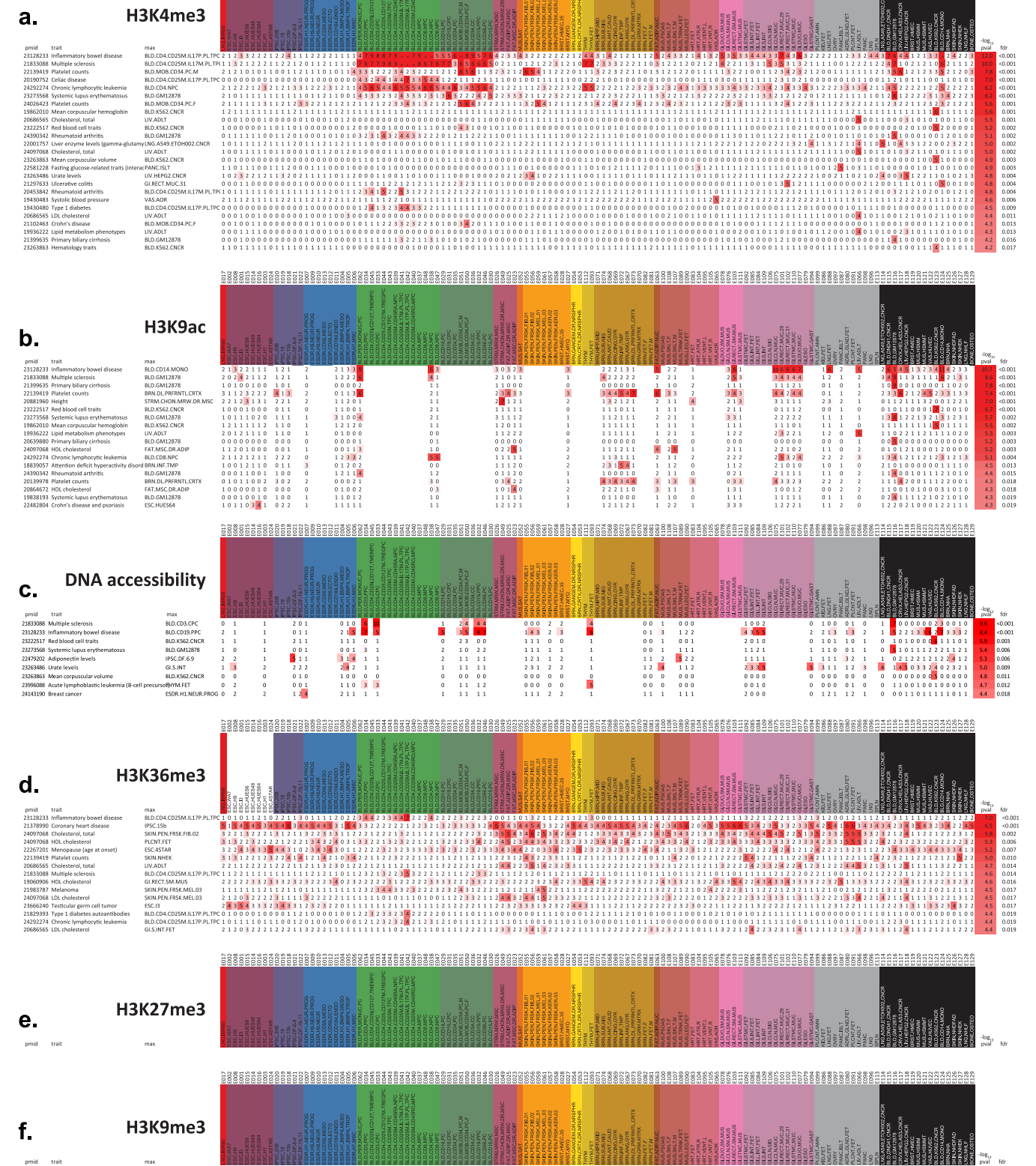
Extended Data Figure 10 | Positional biases of predicted driver motifs relative to high-resolution DNase footprint centres and boundaries.
a, Driver transcription factor motif instance logo, as in Fig. 8 and Extended Data Fig. 9a. **b**, Distribution of motif instances relative to the centre of the high-resolution DNase sites (DGF lengths range from 6 bp to 40 bp), each curve coloured according to the cell/tissue type (from Fig. 2 and Supplementary

Table 5b). **c**, Distribution of shuffled motifs that match composition and number of conserved occurrences in the genome^{70,73}. **d**, Positional bias relative to boundary of DGF region for true motifs, similar to **b**. **e**, Positional bias relative to boundary of DGF region for shuffled motifs, similar to **c**. **f**, Cell types showing significant positional bias after multiple testing correction, coloured according to Fig. 2 and Supplementary Table 5b.



Extended Data Figure 11 | Epigenomic enrichments of genetic variants associated with diverse traits. Tissue-specific enrichments for peaks of epigenomic marks for genetic variants associated with complex disease, expanding Fig. 9. Enrichments are shown for: **a**, H3K4me1 peaks (enhancers). This panel includes all the data shown in Fig. 9, but expands the enrichments shown to all reference epigenomes (columns) for studies (rows) that met the

FDR = 0.02 threshold. **b**, H3K27ac peaks (active enhancers). **a, b**, Studies were defined by a set of SNPs annotated in the GWAS catalogue with the same combination of a publication (shown by the Pubmed ID) and trait. Epigenome with maximum enrichment, uncorrected $-\log_{10} P$ value and estimated FDR are indicated.



Extended Data Figure 12 | Epigenomic enrichments of genetic variants associated with diverse traits. Tissue-specific enrichments for peaks of epigenomic marks for genetic variants associated with complex disease, similar to Extended Data Fig. 11 except enrichments are shown for: **a**, H3K4me3 peaks (promoters); **b**, H3K9ac peaks (active promoters and active enhancers); **c**, DNase peaks (accessible regions); **d**, H3K36me3 peaks (transcribed regions); **e**, **f**, H3K27me3 peaks (Polycomb-repressed regions, **e**) and H3K9me3 peaks (heterochromatin regions, **f**) do not show any enrichments at the FDR = 0.02 threshold. As for Extended Data Fig. 11, studies were defined by a set of SNPs annotated in the GWAS catalogue with the same combination of a trait (far left column) and publication shown by the PubMed ID (far right column), uncorrected P value (in $-\log_{10}$) and estimated FDR.

Chromatin architecture reorganization during stem cell differentiation

Jesse R. Dixon^{1,2*}, Inkyung Jung^{1*}, Siddarth Selvaraj^{1,3*}, Yin Shen¹, Jessica E. Antosiewicz-Bourget⁴, Ah Young Lee¹, Zhen Ye¹, Audrey Kim¹, Nisha Rajagopal¹, Wei Xie⁵, Yarui Diao¹, Jing Liang⁶, Huimin Zhao⁶, Victor V. Lobanenko⁷, Joseph R. Ecker⁸, James A. Thomson^{4,9,10} & Bing Ren^{1,11}

Higher-order chromatin structure is emerging as an important regulator of gene expression. Although dynamic chromatin structures have been identified in the genome, the full scope of chromatin dynamics during mammalian development and lineage specification remains to be determined. By mapping genome-wide chromatin interactions in human embryonic stem (ES) cells and four human ES-cell-derived lineages, we uncover extensive chromatin reorganization during lineage specification. We observe that although self-associating chromatin domains are stable during differentiation, chromatin interactions both within and between domains change in a striking manner, altering 36% of active and inactive chromosomal compartments throughout the genome. By integrating chromatin interaction maps with haplotype-resolved epigenome and transcriptome data sets, we find widespread allelic bias in gene expression correlated with allele-biased chromatin states of linked promoters and distal enhancers. Our results therefore provide a global view of chromatin dynamics and a resource for studying long-range control of gene expression in distinct human cell lineages.

Three-dimensional genome organization is increasingly considered an important regulator of gene expression^{1–4}. Recent high-throughput studies of chromatin structure have begun to shed light on the global organization of our genome^{4–10}. For instance, we and others recently discovered that inter-phase chromosomes are partitioned into megabase-sized topological domains and smaller sub-domains (also known as topologically associated domains or TADs)^{6–9}. These TADs form the basis for higher-level structures referred to as the ‘A’ and ‘B’ compartments^{5,6}. The A and B compartments are closely linked to other functional partitions of the genome, such as early or late DNA replication timing and nuclear lamina association^{11,12}. Despite these advances, our understanding of the dynamic nature of chromatin architecture across human cell types and its effect on cellular identity is incomplete. Here we analyse genome-wide higher-order chromatin interactions in H1 human ES cells and four human ES-cell-derived lineages, mesendoderm (ME), mesenchymal stem (MS) cells, neural progenitor (NP) cells and trophoblast-like (TB) cells¹³. These lineages represent extra-embryonic and embryonic lineages at early stages of development and have been extensively characterized by the Epigenome Roadmap project¹³, with data sets including mRNA-seq, ChIP-seq for 13–24 histone modifications, base-resolution methylC-seq and DNaseI hypersensitivity (DHS) in each lineage^{13,14}. As such, this experimental system provides an opportunity to compare variability in higher-order chromatin structure with underlying gene expression and chromatin state in a genome-wide manner. Further, using a newly developed method to phase two parental alleles into chromosome-span haplotypes from high-resolution chromosome conformation capture (Hi-C) data¹⁵, we have phased the H1 genome to allow for analysis of allele-specific activity



EPIGENOME ROADMAP

A Nature special issue
nature.com/epigenomeroadmap

and chromatin structure. This represents the most extensive data set generated to date, to our knowledge, for the analysis of higher-order chromatin structure, allele-specific chromatin structure and state, and allele-specific gene expression.

Data generation and validation

We performed Hi-C experiments⁵ in two biological replicates in H1 human ES cells and each of the four H1-derived lineages, generating a total of 3.85-billion unique read pairs (Supplementary Table 1). We normalized the intrinsic biases in Hi-C data¹⁶, and confirmed the high reproducibility and accuracy of our Hi-C data sets using several metrics (Extended Data Fig. 1a–d, Supplementary Information and Supplementary Table 2).

Extensive A/B compartment switching

Hi-C interaction maps provide information on multiple hierarchical levels of genome organization⁴. Previous studies demonstrated that the genome is organized into A and B compartments, containing relatively active and inactive regions, respectively^{5,11}. Currently, it is unclear if the A and B compartments change during differentiation and how this relates to lineage specification. We observe a large degree of spatial plasticity in the arrangement of the A/B compartments across cell types, with 36% of the genome switching compartments in at least one of the lineages analysed (Methods; Fig. 1a and Extended Data Fig. 2a–c). Many of the A/B compartment transitions are lineage-restricted (Fig. 1b). Notably, there appears to be a large expansion of the B compartment upon differentiation of human ES cells to MS cells or in IMR90 fibroblasts. These two cell types have previously been shown to undergo an expansion of repressive

¹Ludwig Institute for Cancer Research, 9500 Gilman Drive, La Jolla, California 92093-0653, USA. ²Medical Scientist Training Program, University of California, San Diego, 9500 Gilman Drive, La Jolla, California 92093, USA. ³Bioinformatics and Systems Biology Graduate Program, University of California, San Diego, 9500 Gilman Drive, La Jolla, California 92093, USA. ⁴The Morgridge Institute for Research, 309 North Orchard Street, Madison, Wisconsin 53715, USA. ⁵Tsinghua University–Peking University Center for Life Sciences, School of Life Sciences, Tsinghua University, Beijing 100084, China. ⁶Department of Chemical and Biomolecular Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA. ⁷Laboratory of Immunogenetics, National Institute of Allergy and Infectious Diseases, Twinbrook I NIAID Facility, Room 1417, 5640 Fishers Lane, Rockville, Maryland 20852, USA. ⁸Howard Hughes Medical Institute, The Salk Institute for Biological Studies, 10010 North Torrey Pines Road, La Jolla, California 92037, USA. ⁹Department of Cell and Regenerative Biology, University of Wisconsin School of Medicine and Public Health, Madison, Wisconsin 53706, USA. ¹⁰Department of Molecular, Cellular, and Developmental Biology, University of California Santa Barbara, Santa Barbara, California 93106, USA. ¹¹University of California, San Diego School of Medicine, Department of Cellular and Molecular Medicine, Institute of Genomic Medicine, 9500 Gilman Drive, La Jolla, California 92093-0653, USA.

*These authors contributed equally to this work.

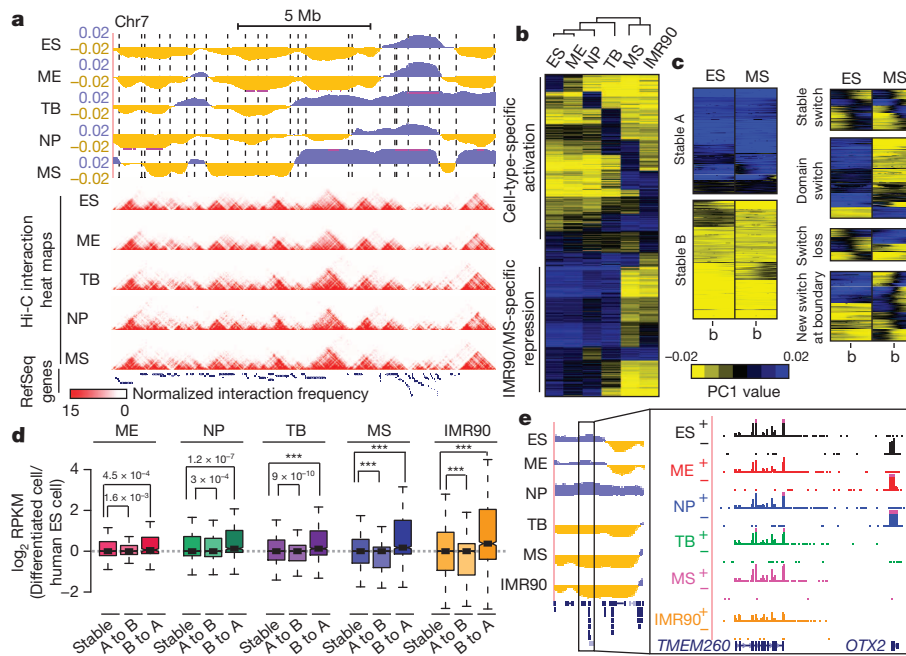


Figure 1 | Dynamic reorganization of chromatin structure during differentiation of human ES cells. **a**, First principal component (PC1) values and Hi-C interaction heat maps in H1 ES cells and H1-derived lineages. PC1 values are used to determine the A/B compartment status of a given region, where positive PC1 values represent A compartment regions (blue), and negative values represent B compartment regions (yellow). Dashed lines indicate TAD boundaries in ES cells. **b**, K-means clustering ($k = 20$) of PC1 values for 40-kb regions of the genome that change A/B compartment status in

at least one lineage. **c**, K-means clustering of PC1 values surrounding TAD boundaries ('b' denotes boundary location). **d**, Distribution of fold-change in gene expression for genes that change compartment status ('A to B' or 'B to A') or that remain the same ('stable') upon differentiation ($***P < 2.2 \times 10^{-16}$, P values by Wilcoxon test; whiskers correspond to interquartile range). **e**, Genome browser for two genes of which one (*OTX2*) shows concordance between expression and PC1 values, whereas a second (*TMEM260*) does not.

heterochromatin modifications during differentiation^{13,17}. In this regard, there appears to be a similar redistribution of the spatial organization of their genomes as well. We observe that the regions that change their A/B compartment status typically correspond to a single or series of TADs (Fig. 1a, c and Extended Data Fig. 2d, e), suggesting that TADs are the units of dynamic alterations in chromosome compartments. Consistent with previous studies of individual loci^{18–20}, we found that genes that change from compartment A to B tend to show reduced expression, whereas genes that change from B to A tend to show higher expression (Fig. 1d). In addition, lineage-restricted compartment A regions tend to include more lineage-restricted genes compared to other regions (Extended Data Fig. 3a). Although statistically significant, the overall patterns of change in expression are subtle. Reasoning that this modest correlation may be due to the possibility that only a subset of genes may be affected by compartment changes, although most genes remain unaffected, we identified a subset of 718 genes with co-variation between gene expression and compartment switching (Fig. 1e, Extended Data Fig. 3b, c, and Methods). These genes were enriched for low CpG content promoters (21.8% versus 15.6% for non-concordant genes, P value 8×10^{-11} , Fisher's exact test), and several significant Gene Ontology (GO) terms, most notably related to extracellular proteins and extracellular matrix (Supplementary Table 3). Taken together, these results indicate that at a global level, there is a high degree of plasticity in the A and B compartments, yet relatively subtle corresponding changes in gene expression, indicating that the A and B compartments have a contributory but not deterministic role in determining cell-type-specific patterns of gene expression.

Domain-level chromatin dynamics

We next examined higher-order chromatin structure at a sub-chromosomal scale. Previous studies indicated that chromosomes are composed of cell-type-invariant TADs^{6,8}. Across the six lineages analysed in this study, we observe that although the positioning of TADs remains stable between cell

types (Fig. 2a), numerous changes in chromatin structure occur within domains. We observed a phenomenon that within some domains, a large portion of the interactions appears to increase or decrease across the entire domain between cell types (Fig. 2b). This suggests that a subset of TADs in a given lineage undergo concerted, domain-wide changes in interaction frequency. Hundreds of TADs underwent such alterations in each lineage (Fig. 2b and Extended Data Fig. 3d), with the changes in interaction frequency correlated positively with active marks such as DHS, H3K27ac and with CTCF binding, and negatively correlated with repressive chromatin modifications such as H3K27me3 and H3K9me3 (Fig. 2c, see Methods for details). TADs that have a concerted increase in intra-domain interaction frequency tend to shift from the B to A compartments, while domains that have a concerted decrease in interaction frequency tend to shift from A to B (Extended Data Fig. 3e, f). Consistent with the changes in chromatin state activity, genes within domains that have increased intra-domain interaction frequency tend to be upregulated, while genes within domains that decrease intra-domain interaction frequency tend to be downregulated (Extended Data Fig. 3g, h).

Chromatin state and dynamic interactions

In order to understand the relationship between chromatin dynamics and other genomic and epigenomic features, we performed integrative analysis of the Hi-C data along with the histone modifications, DHS, and CTCF binding data in the six lineages. Specifically, we asked if particular chromatin state patterns predict changes in chromatin interaction frequency. We divided the genome into 40-kb bins and computed changes in chromatin features in each bin upon differentiation. We then built a Random Forest classification model based on chromatin features to classify local interacting bins as having either increased or decreased interaction frequency (see Methods for details). The model was able to classify regions of the genome that increased or decreased interaction frequency with 73% accuracy (Fig. 2d, 100% graph; Extended Data Fig. 4a), which increased to over 80% when we consider only the highest confidence

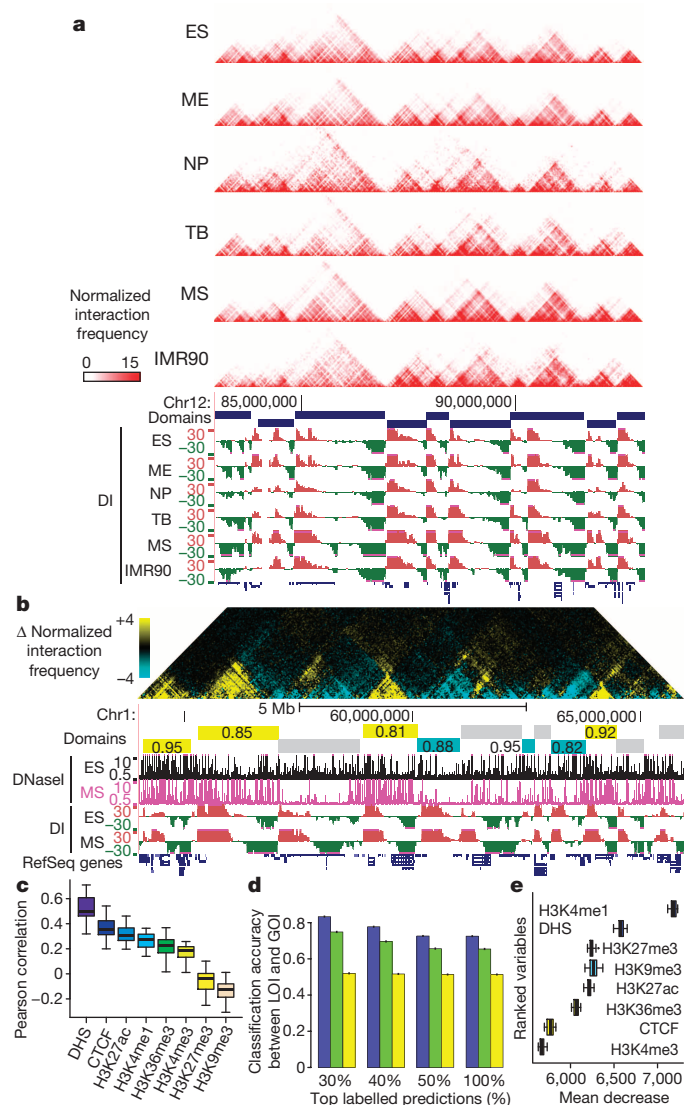


Figure 2 | Domain-wide alterations in chromatin interaction frequency and chromatin state. **a**, Chromatin interaction heat maps in H1 lineages and IMR90 fibroblasts. Also shown are domain calls in ES cells and the directionality index (DI) in each lineage. **b**, Changes in interaction frequency between ES and MS cells. Regions with higher interaction frequency in ES cells are shown in blue, while regions with higher interaction frequency in MS cells are shown in yellow. TADs having a concerted increase or decrease in intra-domain interaction frequency are labelled yellow or blue, respectively, with the fraction of the domain showing increased or decreased interaction frequency listed. Domains that do not show a concerted change are shown in grey. **c**, Boxplots of Pearson correlations coefficients between interaction frequency changes and chromatin mark changes across TADs for each chromosome ($n = 23$). Whiskers correspond to the highest and lowest points within $1.5\times$ the interquartile range. **d**, Classification accuracy of the Random Forest model in predicting whether a bin increases or decreases in interaction frequency ($n = 768,793$), tested on 10 randomly selected subsets of Hi-C data. Accuracy was also checked using actual data (blue), circularized permutation (green) and a random permutation (yellow) of the data. As expected, randomly permuting the data yields 50% accuracy. Accuracy was also assessed considering the top 30, 40, 50% or all predictions based on vote frequency difference (error bars show the standard deviation of accuracies from the 10 randomly selected data subsets). **e**, Ranked chromatin features shown according to importance in classification as boxplots of the mean decrease in Gini index from 10 randomly selected data subsets. Whiskers correspond to the highest and lowest points within $1.5\times$ the interquartile range.

predictions as based on the vote frequency difference (Fig. 2d, 30% graph). The Random Forest model not only indicates that chromatin state features provide information on changes in interaction frequency, it also allows us to determine which chromatin marks are most predictive. Specifically, the 'mean decrease' of the Gini index for each chromatin mark indicates the importance of a given feature during classification. In this regard, we found that change in H3K4me1 density is the most important feature in predicting changes in long-range chromatin interactions (Fig. 2e and Extended Data Fig. 4b, c). As H3K4me1 is present mostly at poised or active enhancers^{21,22}, and as enhancers are known to engage in looping interactions that exist in a cell-type-specific manner²³, these results suggest that enhancer dynamics may play a role in regulating local interaction changes during lineage specification. Consistent with this hypothesis, 40-kb regions with increased interaction frequency tend to have increased enhancer density (Extended Data Fig. 4d, e).

Allele-specific chromatin organization

Normal diploid human cells contain two copies of each chromosome. The collection of variants on a given parental chromosome (also known as the parental haplotype) can be used to determine functional differences between two homologous chromosomes. Previous studies have revealed substantial differences between alleles in gene expression, DNA methylation, and chromatin states^{24–29}. Apart from studies of individual loci in the genome^{30–32}, little is known about the variability in higher-order chromatin structure between homologous chromosomes. Recent work from our laboratory¹⁵ has demonstrated that Hi-C data can be re-purposed to reconstruct chromosome-span haplotypes, which allows for the study of chromatin state and gene expression as a true diploid. We generated chromosome-span haplotypes incorporating $\sim 93.5\%$ of all heterozygous variants for H1 from a combination of Hi-C data sets, whole genome sequencing, and local conditional phasing¹⁵ (Fig. 3a). We observe a high level of concordance among the predicted haplotypes and paired sequence reads from data sets with 'long insert' sizes (Extended Data Fig. 5a), indicating that the reconstructed haplotypes are of high quality. Next, we re-analysed data sets from Hi-C, mRNA-seq, ChIP-seq, methylC-seq, and DNase-seq experiments and determined from which parental haplotype each sequence read was derived (arbitrarily termed the 'p1' and 'p2' allele, as we cannot determine which is the maternal or paternal copy from sequence information alone) (Fig. 3b and Extended Data Fig. 5b).

From the haplotype-resolved A and B compartment patterns across the p1 and p2 alleles in each lineage, we found that homologous chromosomes have highly similar A/B compartment patterns (Fig. 3c and Extended Data Fig. 5c–e), with only 0.6–2.3% of the genome having different A/B compartments between alleles in any given cell type (Extended Data Fig. 5f). Notably, rare regions of the genome do show changes in A/B compartment status between alleles (Fig. 3d), but are not enriched for either allele-biased or known imprinted genes (Extended Data Fig. 5g, h). On the contrary, regions of the genome containing allele-biased or imprinted genes have a subtle but statistically significant increase in the variability of A/B compartment scores between alleles (Fig. 3e). Likewise, the genomic regions with allelic chromatin states have greater variability in A/B compartment scores (Fig. 3f). This indicates that although most allele-biased and imprinted genes do not have differential compartment status between alleles, there may be subtle differences in higher-order chromatin structure between homologous chromosomes at allele-biased regions, reflecting their underlying allele biases in activity. Lastly, similar to A/B compartment patterns, topological domain patterns appear consistent between alleles (Extended Data Fig. 6a, b). Together, these results suggest that the global folding patterns of homologous chromosomes are highly similar.

Allelic imbalances in gene expression

Previous studies of allele-resolved gene expression have identified widespread imbalances in gene expression between different alleles^{24–27,33}. However, it remains unclear to what degree allele-biased gene expression

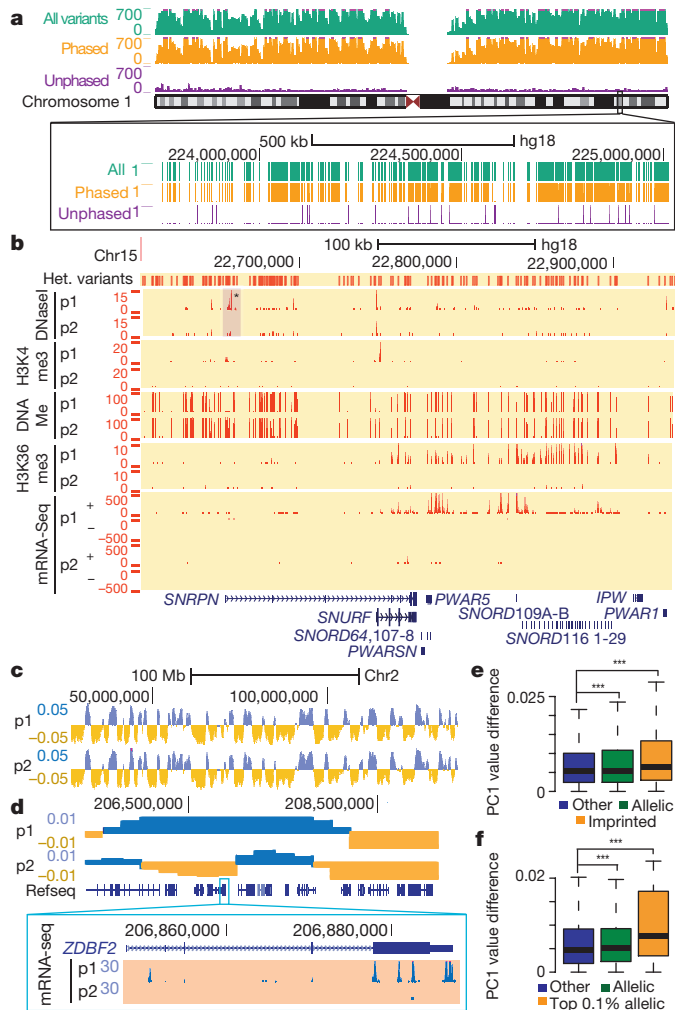


Figure 3 | Haplotype-resolved chromatin organization in H1 lineages.

a, Variants per megabase for all (green), phased (orange) and unphased variants (purple) along chromosome 1. The inset zooms in on a ~1 Mb region, where the presence of a variant at each base is indicated by a value of 1. **b**, Genome browser image of allele specific chromatin features and strand-specific mRNA sequencing. **c**, Genome browser image of PC1 values along chromosome 2 for the p1 and p2 allele. **d**, Allele specific compartment A/B patterns and mRNA-seq surrounding the imprinted *ZDBF2* gene. **e**, Boxplots of the difference between alleles of PC1 values. Regions with imprinted genes ($P = 0.003$) and allelic genes ($P = 0.002$) have more variable PC1 values (Kolmogorov-Smirnov (KS) test). Whiskers correspond to the highest and lowest points within $1.5 \times$ the interquartile range. **f**, Similar to **e**, but for regions with differential allelic chromatin activity (the number of allelic biased variants per 200-kb bin). Regions in the top 0.1% of differential allelic activities (orange) show greater differences in PC1 values compared other regions ($P = 1.6 \times 10^{-8}$ and $P = 0.0015$, respectively, KS test).

varies among different lineages of a single individual. To address this, we re-analysed haplotype-resolved mRNA-seq data and identified allelic biases in gene expression across the five H1 lineages. A total of 1,787 genes showed allelic bias in gene expression in one or more lineages studied here, representing ~24% of all testable genes (false discovery rate (FDR) 10%, Fig. 4a). Most allelic differences in expression are not 'on/off' events, but instead reflect biases in the level of expression from each allele (Fig. 4b). Further, allele-biased genes include both lineage-specific and constitutively expressed genes (Extended Data Fig. 6c, d), and patterns of allelic bias can also be constitutive or cell-type variable (Fig. 4c, d). Only in rare cases do genes switch expression from one allele to the other between cell types.

As expected, genes subject to genomic imprinting are enriched among genes with allelic biases in expression (Fig. 4e), though these represent

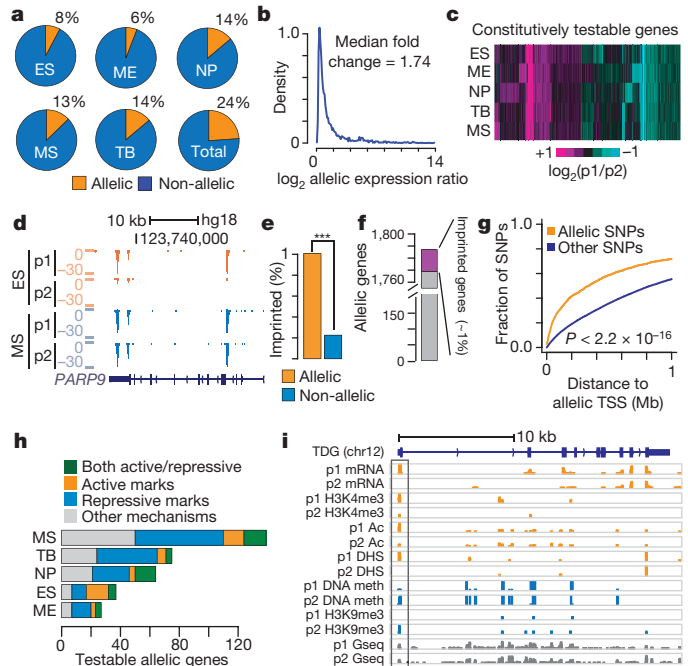


Figure 4 | Allelic biases in gene expression in H1 lineages. **a**, Proportion of genes with detectable allelic expression with statistically significant allelic bias. **b**, Density plot of the absolute value of the fold change in expression (\log_2) between alleles. **c**, Heat map showing k -means ($k = 20$) clustering of the allelic expression ratios (\log_2) at genes with constitutively testable expression (a minimum of 10 reads in each lineage). **d**, Genome browser image of variable allelic expression of the *PARP9* gene. **e**, Fraction of imprinted genes among allele-biased genes and other genes. ($P = 4.4 \times 10^{-5}$, Fisher's exact test). **f**, Fraction of allele-biased genes that are known imprinted genes. **g**, Cumulative density plot of distances from variants to the nearest allele-specific gene. Allele specific variants are defined using histone acetylation, H3K9me3, H3K27me3, DHS and H3K4me3 ($n = 3,920$, $P < 2.2 \times 10^{-16}$, KS test). **h**, Number of allele-biased genes showing consistent allele specific chromatin states in their promoter regions. Active variants are defined by H3K4me3, DHS or histone acetylation. Inactive promoter variants are defined by DNA methylation and H3K9me3/27me3. **i**, Genome browser image of mRNA-seq and chromatin features surrounding the *TDG* gene.

~1% of allele-biased genes (Fig. 4f). Although imprinted genes often occur in clusters, the majority of allele-biased gene expression is not clustered in the genome (Extended Data Fig. 6e). Taken together, these data suggest that most instances of allele-biased gene expression are due to mechanisms other than genomic imprinting. One possible regulatory mechanism that could give rise to allele-biased expression would be allelic bias in activity of *cis*-regulatory elements near these genes. Indeed, regions of the genome that show allelic bias in histone acetylation, histone methylation, CTCF binding, and DHS are closer to allele-biased genes than randomly selected genomic regions (Fig. 4g). Furthermore, allelic gene expression is strongly correlated with DNA methylation or chromatin modification state at promoters (Fig. 4h, i). Of the 247 genes that contain heterozygous variants in their promoter regions and display biased transcription in at least one lineage, a majority exhibit allele-biased chromatin modifications or DNA methylation at the promoter (Fig. 4h). Interestingly, 29% of the testable genes that have allele-biased expression show no evidence of allelic bias in chromatin state or DNA methylation at the promoter (Fig. 4h), raising the possibility that elements outside of promoters may be responsible for the allelic gene expression.

We identified 726, 969, and 5,769 allelic enhancers¹³ that showed allele bias in histone acetylation, DHS, and DNA methylation, respectively (Fig. 5a). We observed a general concordance in allelic biases between enhancers exhibiting allelic histone acetylation and enhancers showing allelic DHS (Fig. 5a). However, we observe only modest concordance between DHS or acetylation defined enhancers with those identified

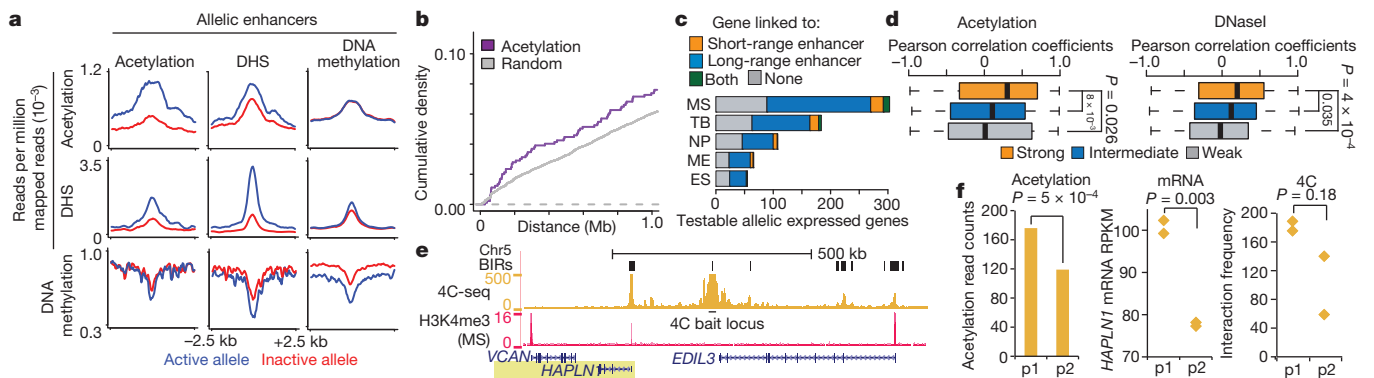


Figure 5 | Allele biases at enhancers in H1 lineages. **a**, Enrichment of acetylation (top row), DHS (middle) and DNA methylation (bottom) at enhancers defined as allelic by acetylation (left column), DHS (middle), or DNA methylation (right). The active allele is in blue, inactive allele in red. **b**, The distance between allelic genes and enhancers as defined by allelic histone acetylation (purple) compared with randomly selected enhancers (grey). Random enhancers were selected to match the read coverage of allele-biased enhancers. **c**, Number of allele specific genes linked to concordantly biased allele specific enhancers. Genes linked by 'long-range enhancers' are defined using Hi-C interaction frequencies, whereas 'short-range enhancers' are defined as any enhancer less than 20 kb from a genes transcription start site. **d**, Boxplots of the Pearson correlation coefficients between allelic gene-enhancer pairs defined by acetylation (left, $n = 1,388$) or DHS (right,

$n = 1,601$). Gene-enhancer pairs are grouped into strongly interacting (top 30%), weakly interacting (bottom 30%), and intermediately interacting pairs (others) based on Hi-C interaction frequency (P values using Welch's t -test). Whiskers correspond to the highest and lowest points within $1.5\times$ the interquartile range. **e**, Normalized 4C-seq interaction frequencies near the *HAPLN1* gene. The 4C-seq bait region is in an allele-biased enhancer near the 3' end of the *EDIL3* gene. Specific interactions called by the LOWESS regression model are shown in black as 'bait interacting regions' (BIRs). **f**, Allele-biased expression of the two alleles of the *HAPLN1* gene, histone acetylation levels at the nearby interacting allele-biased enhancer and allele resolved 4C-seq data (4C-seq P value from t -test, $n = 2$ for p1 allele, $n = 2$ for p2 allele).

based on allelic DNA methylation (Fig. 5a). This may reflect greater power in identifying differentially methylated regions between the two alleles. Alternatively, this may reflect the presence of 'poised' enhancers, where there is not a strict relationship between differences in DNA methylation and enhancer or DHS state^{34,35}. Enhancers with allele-biased acetylation are generally located closer to genes that also show allele-biased expression when compared with enhancers that lack allele bias (Fig. 5b and Extended Data Fig. 6f). A majority (66%) of the 640 allelic genes that display strong Hi-C interactions with allelic enhancers also show concordant allelic activity between the enhancer and promoter (Fig. 5c, Extended Data Fig. 7, and Methods). Additionally, enhancer-gene pairs linked by relatively strong Hi-C interactions show greater correlation between allelic enhancer activity and allelic gene expression compared with pairs linked by weaker Hi-C interactions (Fig. 5d). To test if allelic enhancers indeed form specific contacts with allele-biased genes, we performed 4C-seq^{31,36} with 6 allele-biased enhancers and identified that 4 out of these 6 allelic enhancers showed specific 4C interactions with a nearby allele-biased gene (Fig. 5e, Extended Data Fig. 8 and Supplementary Table 4). Taken together, our results strongly support that allele-biased enhancer activity is a possible mechanism underlying allele-biased gene expression.

To determine if part of the mechanism of regulation by allele-biased enhancers also involved allelic chromatin looping between distal enhancers and putative target genes, we tested for the presence of allele-biased Hi-C reads at allele-biased enhancers throughout the H1 genome by aggregating all Hi-C reads between allelic enhancers and the promoters of nearby allelic genes. We observed that alleles containing enhancer activity generally have higher numbers of chromatin interactions with the target promoters (Extended Data Fig. 9a). This result is confirmed by re-analysis of previous high-resolution 4C-seq results³¹. Two loci (*HAPLN1* and *MAN1C1*) show a similar trend between allele bias in enhancer-promoter interactions with the allelic enhancer acetylation and gene expression levels (Fig. 5f and Extended Data Fig. 9), though the trend in the allelic 4C-seq does not meet statistical significance. The remaining two loci (*FAM65B*, *PXK*) appear to have nearly equal interaction frequencies with the target promoters. Taken together, these results suggest that the allele-biased enhancers can impart allele-biased gene expression either through stable higher-order DNA looping between the two alleles or through potential allele-specific enhancer-promoter interactions.

Discussion

We have presented genome-wide chromatin interaction maps in H1 human ES cells and four H1-derived lineages. We observed dynamic reorganization of higher-order chromatin structure during ES cell differentiation at multiple hierarchical scales. We found extensive switching between the A and B compartments during ES cell differentiation, and observed that distinct subsets of genes have concordant A/B compartments status and expression levels. In this regard, these results are similar to what has been seen with nuclear lamina tethering studies^{20,37–39}, where the expression of only a subset of genes is affected by compartment changes, while other genes remain unaffected. Changes in compartment status may influence the accessibility of genomic regions to transcription factors or other regulatory proteins, which may be particularly important for certain subsets of genes.

In addition, we have observed local alterations in chromatin interaction frequency within TADs. These local changes are best predicted by changes in levels of H3K4me1 and the density of enhancer elements. This is in agreement with recent 5C studies demonstrating that cell-type specific interaction regions are enriched for Smc1, mediator, and transcription factor binding sites⁷. Taken together, these results suggest that enhancer elements likely play an important role in shaping local higher-order chromatin structure throughout the genome. In addition, by analysing patterns of chromatin interactions on each parental allele, we observe relatively minor global changes in higher-order chromatin structure between alleles.

The chromatin interaction maps generated in this study also allowed the reconstruction of chromosome-span haplotypes for the H1 genome. This data set represents one of the first studies of allele-biased expression across multiple cell types of a single individual, as well as analysis of chromatin state at the linked *cis* regulatory elements. Our data set will serve as a valuable tool for the community to better understand the gene regulatory networks controlling pluripotency and differentiation of human embryonic stem cells.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 25 November 2013; accepted 12 January 2015.

- Smallwood, A. & Ren, B. Genome organization and long-range regulation of gene expression by enhancers. *Curr. Opin. Cell Biol.* **25**, 387–394 (2013).

2. Phillips, J. E. & Corces, V. G. CTCF: master weaver of the genome. *Cell* **137**, 1194–1211 (2009).
3. Lettice, L. A. *et al.* Disruption of a long-range cis-acting regulator for *Shh* causes preaxial polydactyly. *Proc. Natl Acad. Sci. USA* **99**, 7548–7553 (2002).
4. Gorkin, D. U., Leung, D. & Ren, B. The 3D genome in transcriptional regulation and pluripotency. *Cell Stem Cell* **14**, 762–775 (2014).
5. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
6. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
7. Phillips-Cremins, J. E. *et al.* Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell* **153**, 1281–1295 (2013).
8. Nora, E. P. *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381–385 (2012).
9. Sexton, T. *et al.* Three-dimensional folding and functional organization principles of the *Drosophila* genome. *Cell* **148**, 458–472 (2012).
10. Rao, S. S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
11. Ryba, T. *et al.* Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Res.* **20**, 761–770 (2010).
12. Peric-Hupkes, D. *et al.* Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. *Mol. Cell* **38**, 603–613 (2010).
13. Xie, W. *et al.* Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell* **153**, 1134–1148 (2013).
14. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
15. Selvaraj, S., Dixon, J. R., Bansal, V. & Ren, B. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nature Biotechnol.* **31**, 1111–1118 (2013).
16. Hu, M. *et al.* HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics* **28**, 3131–3133 (2012).
17. Hawkins, R. D. *et al.* Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell Stem Cell* **6**, 479–491 (2010).
18. Brown, K. E. *et al.* Association of transcriptionally silent genes with Ikaros complexes at centromeric heterochromatin. *Cell* **91**, 845–854 (1997).
19. Kosak, S. T. *et al.* Subnuclear compartmentalization of immunoglobulin loci during lymphocyte development. *Science* **296**, 158–162 (2002).
20. Holwerda, S. & de Laat, W. Chromatin loops, gene positioning, and gene expression. *Front. Genet.* **3**, 217 (2012).
21. Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genet.* **39**, 311–318 (2007).
22. Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108–112 (2009).
23. Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489**, 109–113 (2012).
24. Heinz, S. *et al.* Effect of natural genetic variation on enhancer selection and function. *Nature* **503**, 487–492 (2013).
25. McVicker, G. *et al.* Identification of genetic variants that affect histone modifications in human cells. *Science* **342**, 747–749 (2013).
26. Kasowski, M. *et al.* Extensive variation in chromatin states across humans. *Science* **342**, 750–752 (2013).
27. Kilpinen, H. *et al.* Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science* **342**, 744–747 (2013).
28. Kuleshov, V. *et al.* Whole-genome haplotyping using long reads and statistical methods. *Nature Biotechnol.* **32**, 261–266 (2014).
29. Xie, W. *et al.* Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome. *Cell* **148**, 816–831 (2012).
30. Splinter, E. *et al.* The inactive X chromosome adopts a unique three-dimensional conformation that is dependent on Xist RNA. *Genes Dev.* **25**, 1371–1383 (2011).
31. Holwerda, S. J. *et al.* Allelic exclusion of the immunoglobulin heavy chain locus is independent of its nuclear localization in mature B cells. *Nucleic Acids Res.* **41**, 6905–6916 (2013).
32. de Wit, E. *et al.* The pluripotent genome in three dimensions is shaped around pluripotency factors. *Nature* **501**, 227–231 (2013).
33. Gimelbrant, A., Hutchinson, J. N., Thompson, B. R. & Chess, A. Widespread monoallelic expression on human autosomes. *Science* **318**, 1136–1140 (2007).
34. Hon, G. C. *et al.* Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues. *Nature Genet.* **45**, 1198–1206 (2013).
35. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).
36. van de Werken, H. J. *et al.* Robust 4C-seq data analysis to screen for regulatory DNA interactions. *Nature Methods* **9**, 969–972 (2012).
37. Reddy, K. L., Zullo, J. M., Bertolino, E. & Singh, H. Transcriptional repression mediated by repositioning of genes to the nuclear lamina. *Nature* **452**, 243–247 (2008).
38. Finlan, L. E. *et al.* Recruitment to the nuclear periphery can alter expression of genes in human cells. *PLoS Genet.* **4**, e1000039 (2008).
39. Kumaran, R. I. & Spector, D. L. A genetic locus targeted to the nuclear periphery in living cells maintains its transcriptional competence. *J. Cell Biol.* **180**, 51–65 (2008).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank the members of the Ren laboratory for support and suggestions throughout the course of this work. This work is funded in part by the Ludwig Institute for Cancer Research, the NIH Roadmap Epigenome Project (U01 ES017166 and R01 ES024984), and the California Institute of Regenerative Medicine (RN2-00905). Y.D. is supported by a postdoctoral fellowship from the Human Frontier Science Program (LT000576/2014-L).

Author Contributions J.R.D., I.J., S.S., and B.R. conceived the study. J.R.D. performed Hi-C and 4C experiments with assistance from A.K. J.R.D., I.J., and S.S. performed the data analysis. A.Y.L. performed ChIP-seq experiments for CTCF with assistance from Z.Y. J.E.A.-B. performed ES cell culture, differentiation and sample collection. N.R. and W.X. provided expertise and assistance in data analysis. Y.S., Y.D., V.V.L. J.L., H.Z., J.R.E., and J.T. provided insights and support for the design and execution of the project. J.R.D. and I.J. prepared the manuscript with assistance from S.S. and B.R.

Author Information All data from this study have been deposited in the GEO database under accession number GSE52457. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to B.R. (biren@ucsd.edu).



This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported licence. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons licence, users will need to obtain permission from the licence holder to reproduce the material. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-sa/3.0>

METHODS

Cell culture and previous data sets analysed. H1 human ES cells and H1-derived cells were cultured as previously described¹³. ChIP-seq experiments for CTCF were performed using previously published methods and antibodies^{13,40}. Hi-C libraries were generated as previously described⁷. Two biological replicates of Hi-C data were generated for each lineage in order to assess the reproducibility of the data. Hi-C and ChIP-seq libraries were sequenced on the Illumina Hi-Seq 2000 and Hi-Seq 2500 platforms. mRNA-seq, ChIP-seq for histone modifications and methylC-seq data sets have been previously published¹³. DNase-seq experiments have been previously described elsewhere¹⁴.

Sequence read alignment. The following description applies for the alignment of DNA methylation, ChIP-seq and DNase-seq data sets. Single-end sequencing data was mapped to a variant masked reference genome (hg18) using Novoalign. Unmapped and non-uniquely mapping reads were removed, and PCR duplicate reads were removed with Picard. Reads were processed with the Genome Analysis Toolkit (GATK)⁴¹. Specifically, reads underwent indel recalibration and variant realignment. Lastly, reads that overlapped with variant loci were split into the 'p1' and 'p2' allele according to whether the base in each sequencing read matched the sequence from either the p1 or the p2 alleles.

For Hi-C data sets, read pairs were mapped independently to the variant masked genome using Novoalign. Reads were then manually paired using in house scripts. Non-uniquely mapping, unmapped reads and PCR duplicate read pairs were removed. Reads pairs were then split into single reads and processed through the same GATK pipeline described above including indel re-alignment and variant recalibration. Finally, read pairs were manually re-paired using in house scripts.

For mRNA-seq, we mapped the paired-end data to a variant masked reference. We used Useq software to first process the variant masked genome to create a splice junction reference. Reads were then mapped to the Useq processed reference genome using Novoalign. Lastly, we converted the read alignment locations from the Useq processed genome back to hg18 coordinates using Useq.

Whole-genome sequencing, genotyping and haplotyping. Whole genome sequencing (WGS) data for the H1 genome was downloaded from the Sequence Read Archive database (SRA049981). Reads were mapped to the hg18 reference genome using Novoalign. Unmapped and non-uniquely mapping reads were removed using in house scripts. PCR duplicate reads were removed using Picard. The data was processed through the Genome Analysis Toolkit (GATK) best practices guidelines. We performed indel recalibration, variant realignment, variant calling using the Unified Genotyper, and variant recalibration.

Haplotyping was performed using the previously described HaploSeq method¹⁵. Briefly, Hi-C reads from each of the H1 derived lineages were used as input sequencing into the HapCUT software⁴² in order to generate haplotype predictions. For final haplotype calls, Hi-C data was combined with WGS mate-pair data for the H1 genome. HapCUT generates several 'blocks' for each chromosome. The vast majority of variants on each chromosome are in the 'most variants phased' (MVP) block. The MVP block for each chromosome was used as a 'seed haplotype' for local conditional phasing using population sequencing data from the 1000 genomes project using the Beagle v.4.0 software⁴³. This generates two haplotypes for each chromosome, one for the maternal allele and one for the paternal allele. As we do not have information regarding the parent of origin in the H1 genome, we arbitrarily define each allele as the p1 or p2 allele (p1 and p2 for parent 1 and 2, respectively). The p1 and p2 allele for different chromosomes are not necessarily derived from the same parent, as this information is only accessible if the sequence of H1's parents were also available.

Haplotype alignment bias. Although we mapped the ChIP-seq, DNase-seq, Hi-C and DNA methylation data sets to a variant masked genome, we recognize that there could still be local alignment biases favouring a given allele. To account for this, we performed a two-step filtering process. First, we generated simulated reads that span each position surrounding a variant location in the genome. SNPs and indels that showed >5% and >10% biases, respectively, were excluded from all downstream analyses, as these variants show an inherent mapping bias. Second, for each variant in the genome, we calculated the coverage over the variant based on WGS data. Based on the WGS data, we expect each variant to have near equal coverage between the two alleles. Any variant that had sequencing coverage greater than 3 standard deviations above the mean for each haplotype along a chromosome was excluded, as were variants that showed a Benjamini corrected binomial *P* value of ≤ 0.05 when comparing the WGS read coverage on each allele. Lastly, analysis of allele-biased coverage at a SNP level can be very sensitive to genotyping errors, in particular if a homozygous variant is erroneously called as heterozygous. To account for this we made a null hypothesis that all called heterozygous variants were actually homozygous. We excluded any heterozygous variant with a GATK derived genotype *P* value of greater than 0.05 (after Benjamini correction). This excluded roughly 2% of all heterozygous SNPs in the genome as having genome sequencing coverage that could be expected for a homozygous variant.

Estimation of random collision events in Hi-C data. We estimated random collision events by calculating the intermolecular ligation rate between a nuclear chromosome (chrN) and the mitochondrial chromosome (chrM). The interacting space between chrN and chrM can be defined by multiplying (roughly 16 kb per chrM \times number of chrM per chrN) and (roughly 6.16 Gb per diploid nucleus). The number of chrM per chrN was calculated from ChIP-seq input sequencing data.

Number of chrM per chrN = Number of read counts for chrM/number of read counts for chrN \times 6.16 Gb per chrN \times 16 kb per chrM.

The number of random collision events between any given two loci (40-kb bin size) was estimated as following.

Number of random collision events per 40 kb² = number of intermolecular interactions between chrN and chrM/interacting space between chrN and chrM \times 40 kb².

The estimated random collision events are summarized in Supplementary Table 2.

Topological domain calling. We systemically identified topological domains based on the directionality index (DI) score and a Hidden Markov Model (HMM) as previously described⁶. The number of identified topological domains across human genome was 2,468, 2,489, 2,202, 2,144 and 2,407 for ES, ME, MS, NP and TB cells, respectively. According to the topological domain patterns, genomes were partitioned into domains, boundaries and unstructured regions as previously described.

Identification of A and B compartments. Identification of A and B compartments was performed conceptually similarly to what has been previously described⁵, though with several modifications. We used the normalized 40-kb interaction matrices for each cell type and calculated the expected interaction frequency between two 40-kb bins given the distance separating them in the genome. We used a sliding window approach with a bin size of 400 kb and a step size of 40-kb to generate an observed/expected matrix. The observed frequency was the sum of all observed interaction frequencies of the 40-kb bins making up the larger 400-kb bin. Likewise, the expected frequency was the sum of the expected frequencies of each of the 40-kb bins making up the larger 400 kb bin. This value was used to generate the observed/expected. This was then converted to a Pearson correlation matrix and subsequently used for principal component analysis as previously described⁵. Specifically, we used the 'cov' function in R to generate a covariance matrix from the Pearson correlation matrix, and then we used the 'eigen' function in R to generate Eigen vectors and Eigen values from the covariance matrix. The first principal component for each chromosome was used to identify regions of the genome as belonging to either the A or B compartment. The direction of the Eigen values is arbitrary, and therefore positive values were set to 'A' and negative values were set to 'B' based on their association with gene density.

To identify regions of the genome that switched A/B compartment status with differentiation, we first identified regions with statistically significant variability in PC1 values across all cell types using ANOVA. Second, we considered only regions where both biological replicates showed changes in PC1 values from positive to negative or vice versa. This allowed us to define the 36% of the genome that changes compartment status in at least one lineage.

Identification of genes with concordant expression and A/B compartment status.

To define genes with concordant changes in expression and compartment status, we calculated the covariance between the vector of the log₂ of gene expression values and vector of PC1 values for each gene across the six lineages analysed. We use this calculated covariance as a metric to quantitatively define 'concordance'. To calculate a *P* value for the covariance for each gene, we compared these observed covariance values to a random background distribution. The background distribution was generated by randomly shuffling the vector of log₂ of gene expression for each gene and then calculating the covariance between the random gene expression vector and the PC1 values. This was repeated 1,000 times for each gene, and a rank-based *P* value could then be calculated for the observed covariance values. These genes were shown to be enriched for low CpG content promoters, which is defined here by an observed/expected CpG content of <0.35. GO terms analysis of this subset of genes was performed using the DAVID GO terms website.

Identification of A and B compartments in each allele. Identification of A and B compartments in each allele was performed similarly as described in the above section, though with several modifications. Due to the low density of Hi-C interaction frequencies in each allele, we used a sliding window approach with a bin size of 1-Mb and a step size of 200-kb to generate an observed/expected matrix. The first principal component in each allele was used to identify regions of the genome as belonging to either the A or B compartment. The direction of the Eigen values is arbitrary, and therefore the direction was determined according to the correlation coefficient values with the PC1 values generated in the above section.

Changes in intra-domain interaction frequency. To compute the change in interaction frequency between cell types, we first merged the Hi-C data between two replicates for each cell type. The merged, normalized interaction matrices were quantile normalized between all lineages to accommodate for differences in frequency strictly due to sequencing depth. The differences between cell types were computed

by simply subtracting the interaction frequency of each bin I_{ij} of ES cells from the differentiated cell types (as shown in Fig. 2b).

To assess for concerted domain-wide changes in interaction frequency, we calculated two values for each domain: the fraction of interacting bins in the domain that showed an increase in interaction frequency and the fraction of bins that showed a decrease in interaction frequency. To compare these numbers to what would be expected at random, we calculated the same two values for each domain where the bins of the domain were made up of randomly selected intra-domain interacting bins from throughout the genome, keeping the portion of bins in each domain separated by a given genomic distance constant. This randomization was performed 10,000 times for each domain. At random, each domain on average had roughly 50% of bins that increased in interaction frequency and 50% that decreased in interaction frequency. By seeing deviations from these expected values, we could assess for 'concerted' changes in interaction frequency. We assigned a rank-based P value of the degree of 'concertedness' for each domain by comparing the actual observed portion of the domain that was either increased or decreased in interaction frequency with what was observed at random for each domain. These P values were adjusted for multiple testing using Benjamini correction, and we considered any domain as having undergone a concerted change if the final corrected P value was less than 0.001 (0.1% FDR). **Changes in intra-domain interaction frequency between alleles.** Domain-wide interaction frequency differences between alleles were calculated by using the same approach described in the above section. If the domain-wide average interaction frequency difference between alleles was significantly more than randomized data (P value 0.001), the corresponding domains are considered as having allele specific domain-wide interaction frequency changes.

Correlation coefficient between domain-wide interaction frequency changes and modification changes. The domain-wide correlations shown in Fig. 2c between changes in interaction frequency and various chromatin marks were calculated as follows. For each domain, the intra-domain interaction frequency differences between ES cells and each differentiated lineage was calculated for each 40-kb interacting bin of the domain (where we define a single 'interacting bin' as being formed by the interaction of two underlying 40-kb genomic bins). These values were considered as the first vector for the correlations. The vector of histone modification values was calculated as follows. For each 40-kb interacting bin, the enrichment of a given chromatin mark in the two 40-kb bins that compose the interaction was averaged. The average enrichment was then multiplied by a weight proportional to the genomic distance between the two 40-kb bins. This weight was based on the global average of Hi-C interaction frequencies from six lineages analysed between loci separated by a given genomic distance. The two vectors were used to calculate a Pearson correlation in each chromosome, which reflects how change in domain-wide interaction frequency correlates with domain-wide chromatin mark changes.

The Random Forest classification model. We built a Random Forest model to better understand which chromatin modifications may be most predictive of changes in interaction frequency between any two given loci. The aim of the Random Forest model was to classify 40-kb interacting bins as either increased or decreased in interaction frequency given information about the enrichment of various chromatin marks, DHS and CTCF binding sites. The utility of the Random Forest model is twofold: first, by assessing the accuracy of the model using observed data, we can learn whether the information supplied to the model (in this case the chromatin state, DHS and CTCF data) is predictive of the outcome, namely changes in interaction frequency. The second powerful aspect of the model is that it allows us to assess which input data supplied to the model is most informative, allowing us to determine which chromatin state features may be most predictive of changes in higher-order chromatin structure.

The model was built as follows: 40-kb interacting bins in the genome were classified into two groups, ones that increased in interaction frequency, and ones that decreased in interaction frequency. These changes were defined if the 40-kb based interaction frequencies increased or decreased more than twofold in the differentiated lineage compared to those in H1 ES cells. We only considered interacting bins separated by less than 2 Mb. We added a pseudocount value to the average interaction frequencies when we calculate fold changes to allow for comparison of zero values. The resulting criteria yielded 768,793 interacting bins as either losses or gains. Chromatin state changes of H3K4me1, H3K4me3, H3K9me3, H3K27ac, H3K27me3, H3K36me3, DHS signal, and CTCF were also calculated. For each 40-kb bin, RPKM values for each chromatin mark were calculated. Fold changes of RPKM values were calculated by comparing with RPKM values in H1 ES cells. Those 8 chromatin marks were assigned to each interacting region, thus for each interaction (consisting of two interacting 40-kb bins) we can construct a 1 by 16 feature vector.

Using those feature vectors, we built a classification model between gain and loss of interactions using a Random Forest R package with default parameters except for specifying the model to use 500 trees. The performance was measured according to two criteria, the out-of-back error rate achieved from the Random forest model and the tenfold cross validation. We compared out-of-back error rate to tenfold cross validation and observed very similar results (shown in Extended Data Fig. 4a).

As a final result, the Random Forest model gives vote frequencies for predicting whether a given interaction is increased or decreased. The difference in vote frequency between the two states reflects the confidence of the model in a given prediction, with a larger vote frequency difference indicating a higher degree of confidence. The sum of the vote frequencies is equal to 1. As an example, in the case where the model could not predict any changes in interaction frequency, the vote frequencies would be expected to both equal 0.5. If the vote frequency for the 'loss of interaction' class was greater than 0.5, the interacting bin would be classified as having undergone a loss of interaction. Likewise, if the 'gain of interaction' vote frequency was greater than 0.5, the bin was classified as a gain of interaction. Again, the difference in vote frequencies between the two classes reflects the degree of confidence of the model in a given prediction.

When we built the classification model, the balance for the number of inputs between two classes is important. If the model includes more gain of interaction features rather than loss of interaction features, the model is more likely trained to predict a gain of interactions. To avoid this issue, we randomly selected the same number of gain of interaction and loss of interaction feature vectors while building the classification model.

The Random Forest model also provides a measure of the importance of each variable during classification as the 'mean decrease' metric of the Gini index. For a given variable, higher the mean decrease in Gini index, the more important the variable is during classification.

Identification of allelic biased genes, enhancers and SNPs

Allelic genes. We considered the two replicates of mRNA-seq data and used a negative binomial distribution (10% FDR) to calculate significantly biased genes between the two alleles, where genes are defined by merging isoforms (from RefSeq). We used the edgeR software package in R for calculating the P values.

Allelic SNPs. We estimated if a SNP is allele-biased on different types of readouts. In particular, we used ChIP-seq, DHS, and CTCF data sets independently to obtain readouts of each SNP between the two alleles. We then used a binomial statistic (with an expectation $P = 0.5$) to identify significantly biased SNPs for a given data set. FDR was based on 1,000 random permutations.

Differential methylation among alleles (DMRs). Bisulfite sequencing reads were mapped using Novoalign methylation aligner to an H1 variant masked hg18 reference genome. Duplicated and poorly mapped reads were removed, and the reads that contain SNPs were retained for downstream analyses. Reads were then assigned to either the p1 or the p2 allele on the basis of the SNPs present in each read. During this assignment, certain SNPs could not be resolved between the two alleles because of considerations of bisulfite conversion. Specifically, when a SNP is C/T (or listed as A/G on the reverse strand), the conversion of methyl-C to T by bisulfite will make it impossible to distinguish whether a given read is a methylated cytosine from one allele or a thymidine from the other allele. In these cases, these SNPs were excluded from distinguishing from which allele a given read was derived. After resolving into each allele, CpGs were called and nearby CpG were merged (within 100 bp). Of note, in instances where a SNP contains a cytosine, it would be impossible to distinguish whether a difference between two alleles is due to the polymorphism or due to the change in methylation. As such, any position in the genome with a SNP was excluded from our calculation of the percentage methylation over a given window. We called ASM in each of these CpGs using Fisher's exact test with 10% FDR after multiple testing correction as a threshold for significance. We randomly shuffled the methylation and unmethylation values for a given haplotype (for a CpG) and used these random estimates to obtain FDR.

Allelic enhancers. To study allele bias at enhancers, we first calculated the combined coverage of whole genome sequencing data and bisulfite sequencing (without regard for methylation status). Any enhancer where one of the two alleles contained less than 35% of the total allele resolved reads at the enhancer was excluded as having an inherent bias in mapping between the two alleles. To systematically study allelic enhancers, we combined several enhancer marks to obtain a combined acetylation bam file. This combined bam file gives us the required coverage in an allelic context to perform an in-depth analyses. In particular, we combined data from H4K8ac, H4K91ac, H2BK120ac, H3K18ac, H3K23ac, H3K27ac, H3K4ac, H2AK5ac and H3K9ac marks. Using this combined bam file, we examined allelic SNPs described as above. For evaluating allelic enhancers, we obtained readout for enhancers defined in ref. 13 (± 2.5 kb from enhancer peaks) between the two alleles. Then we used binomial to obtain significance at an FDR of 10%, as evaluated by the random permutation analyses (1,000 permutations). The same analysis was used to call allele-biased enhancers based on DHS data. For the analysis of allele bias in DNA methylation at enhancers, we considered any enhancer as having allele-biased DNA methylation if at least one ASM bin overlapped with the enhancer. If more than one bin of ASM overlapped an enhancer, we checked to see whether the patterns of ASM were concurrent between all bins. If there were divergent patterns between ASM bins at an enhancer, these enhancers were excluded.

Distance of allelic enhancers to allelic genes. We compared the distance between allele-biased enhancers, as identified by histone acetylation levels with randomly

selected enhancers, to test the hypothesis that if allele-biased enhancers regulate allele-biased genes, they should generally be closer to allele-biased genes than should randomly chosen enhancers (Fig. 5b). This analysis was complicated by the fact that the rates of heterozygous SNPs near allele-biased genes are higher than for non-allele-biased genes in the genome (Extended Data Fig. 6f). This creates a situation of possible ascertainment bias, owing to the fact that enhancers near allele-biased genes will therefore tend to have slightly higher allele-resolved read coverage as compared with randomly chosen enhancers throughout the genome. To account for this, when comparing the distance of allele-biased enhancers to allele-biased genes with randomly chosen enhancers, we selected random enhancers to match the coverage profile of allele-biased enhancers. This was accomplished by binning all enhancers into increments of 50 sequencing reads, from 0 to 49, 50 to 99, etc., up to 1,700 reads. For each identified allele-biased enhancer, we selected 100 random enhancers from the same coverage bin. This limits the effects of local variation in heterozygosity rates throughout the genome on the likelihood of identifying allele-biased enhancers near allele-biased genes. As such, the results in Fig. 5b are probably not due to the possibility of having greater statistical power for calling allele-biased enhancers near allele-biased genes (because of greater heterozygosity rates and higher numbers of allele-resolved reads).

Enhancers, gene expression levels, lineage-specific genes, housekeeping genes and imprinting genes. The enhancer regions were defined as previously described⁴⁴. Briefly, enhancer chromatin signatures were trained for p300 binding sites in H1 ES cells using RFECS algorithm based on H3K4me1, H3K4me3 and H3K27ac signals at 100-bp bin size. Next, these modification signals in all cell lines were tested to predict enhancers. The predicted enhancers that overlap with H3K4me3 peaks or within 2.5 kb of the transcription start site were removed. Enhancers were merged from all cell types if they are located close to each other (<2 kb) by taking the midpoint at the centre of the new enhancer.

For the gene list, gene expression levels, housekeeping genes and lineage-specific genes we used the same data set as described in ref. 13. For imprinting genes, we obtained known imprinted genes downloaded from publicly available imprinting gene database (<http://www.geneimprint.com/>).

Linking between allelically expressed genes and allele-biased promoter activities. To investigate how many allele-biased gene promoter activities are consistent with allelic gene expression levels, first we selected allelic genes that contain at least one allelic SNP in their promoter regions (1.5 kb upstream and downstream from transcription start site). We only considered allelic SNPs defined by DHS, H3K4me3, histone acetylation, combined H3K9me3 and H3K27me3, and DNA methylation because the signatures of those chromatin marks at the promoter regions are well defined. If promoters are marked by allelic SNPs from H3K9me3/H3K27me3 or DNA methylation and the allelic gene expression levels are consistent with the allele-biased promoter activities, the genes can be explained by allelic repressive marks. If promoters are marked by allelic SNP from histone acetylations, H3K4me3, and DHS, and allelic gene expression levels are consistent with the allele-biased promoter activities, the genes can be explained by allelic active marks.

Identification of enhancer–promoter interactions. To investigate the linking between allelic genes and allelic enhancers we first defined enhancer–promoter interactions using Hi-C interaction frequency data. Hi-C interaction frequencies were calculated in terms of 5-kb windows and normalized using HiCNorm. After that, we considered all pairs of promoters and enhancers in each chromosome. Promoter regions were fixed as ± 5 kb surrounding transcriptional start sites and enhancer regions were defined by using different window sizes as: 5 kb, 10 kb, 20 kb, 30 kb, 40 kb, 50 kb, 75 kb, 100 kb, 300 kb and 500 kb surrounding the centre of each enhancer (Extended Data Fig. 7). The interaction frequencies between a promoter and an enhancer at a certain window size were calculated as (interaction frequency / window size of an enhancer) \times 5 kb. Final interaction scores were defined as summation of interaction frequencies between promoter and enhancer with multiple window sizes. To calculate significance of each enhancer–promoter interaction, we generated a random interaction frequency score by randomly permuted interaction frequencies between the promoter and enhancer in each window size. The distribution of random interaction frequency scores was fit to Weibull distribution and then *P* values of the interaction frequency in each enhancer–promoter pair were calculated. At a given *P* value cutoff, we defined enhancer–promoter interactions. At a *P* value cutoff of 1×10^{-3} , more than 80% of interactions are reproducible between two biological replicates (Extended Data Fig. 7b). By taking this *P* value cutoff, we defined 339,761, 354,529, 319,169, 158,453, 250,495, and 210,010 enhancer–promoter interactions for ES, ME, MS, NP, TB and IMR90 cell lines between 103,982 enhancers and 18,532 promoter regions. These enhancer–promoter interaction numbers can be changed according to cutoff *P* values.

Comparison of enhancer–promoter interaction with other experiments. To validate predicted enhancer–promoter interactions we compared the interaction frequency scores to 5C scores and DNaseI quantitative trait loci (dsQTL) information. 5C is a chromosome conformation capture (3C)-based approach to measure

the interactions of all versus all targeted regions. For the H1 ES cell lines, we downloaded previously generated 5C data²³ and compared this to our interaction frequency between enhancers and promoters. We observe very strong correlative patterns between 5C and our interaction frequency scores (Extended Data Fig. 1b). We can also observe that interacting pairs tend to show higher 5C scores compared to non-interacting pairs.

We also compared interaction frequency scores to dsQTL relationships. dsQTL data provide functional relationships between DHS and their target promoters based on QTL information⁴⁵. We calculated enhancer–promoter interactions scores again for all pairs of DHS and promoter regions based on dsQTL data. Interactions defined by dsQTL were considered as target relationships, otherwise off-target relationships. According to interaction distance, enhancer–promoter interaction scores were calculated between target and off-target gene relationships. We observe that target gene relationships tend to have higher interaction frequency scores (Extended Data Fig. 7d). **Pearson correlation coefficient between allele-biased gene–enhancer pairs.** We calculated Pearson correlation coefficients between allele-biased gene–enhancer pairs. First we generate 1 by 10 vectors for each allelic gene and allelic enhancer using the data from H1 human ES cells and the four H1-derived lineages. For each lineage, we assigned \log_2 (p2 allele read counts / p1 allele read counts) and \log_2 (p1 allele read counts / p2 allele read counts) values as allelic bias information. After constructing two 1 by 10 vectors for both allelic gene and allelic enhancer, we calculated the Pearson correlation coefficient between them.

Identification of allelic Hi-C interactions. We investigated allelic Hi-C interactions for allelic gene–enhancer pairs. We considered allelic interactions between 10-kb surrounding regions for both the transcription start site and enhancer, respectively. Many of allelic gene–enhancer pairs do not have any allelic interactions, but allelic gene–enhancer pairs show concordance if they are connected by allelic Hi-C interactions.

4C-seq experiments and data analysis. 4C-seq was performed essentially as described previously³⁶. Six bait regions were chosen at allele-biased enhancer elements containing SNPs that would allow for performance of allele specific 4C-seq, as has been previously described³¹. Primers were designed such that the first read of a paired-end read would sequence the primer sequence derived from the bait region and read into the target region of interest. The second read in the pair would read a portion of the bait region containing the SNP of interest (see Extended Data Fig. 8 for a diagram of the experimental strategy). The primers were designed to include the Illumina adaptor sequences necessary for sequencing as well as the presence of barcodes derived from Illumina's TruSeq adaptors that allowed for multiplexing of 4C-seq reactions. We used two 4 base cutters, NlaIII and DpnII, for the first or second restriction enzyme digestion, depending on the locus in question (See Supplementary Table 4). 4C-seq templates were prepared as previously described³⁶. 16 PCR reactions using 200 ng of 4C template were performed for 30 cycles for each bait region and pooled together. The PCR reactions underwent a final purification step using AMPure beads (Beckman-Coulter) according to the manufacturer's instructions (using a bead-to-sample ratio of 1.8). The concentrations of each 4C library were calculated using the KAPA qPCR system using a standard curve. The libraries were then combined and spiked in with other non-4C sequencing libraries for sequencing on the Illumina Hi-Seq 2500 machine.

Sequence reads were processed as follows. For each read, the first and second sequencing reads were checked to identify the presence of the primer sequences and any expected portion of the bait region. Any sequence with greater than 20% mismatches to the expected bait region was discarded. The reads were trimmed such that each read was represented as a 36-mer, with 20 bp derived from the bait region and the subsequent 16 bp, presumably containing the target region of interest. Based on the SNP identified in the second sequencing read derived from the bait region, each of these files were split into allele specific 4C-seq FASTQ files for further analysis.

4C-seq data was mapped to a version of the hg18 genome with known SNPs in the H1 genome masked to N, similar to other the strategy of mapping other sequence read data sets performed in this study. Custom indexes for this H1-masked hg18 genome were built using the 4cseqpipe “-build_re_db” command. The reads were mapped using the 4cseqpipe software “-map” command to custom built indexes. Normalized contact intensities were derived using the 4cseqpipe “-nearcis” command for a 1-Mb region upstream and downstream of the bait locus. We then took the normalized fragment level interaction frequency tables and removed any fragments where a SNP either could create or disrupt a potential restriction enzyme site between the two alleles. In addition, given the short sequencing read length, any fragment with an insertion or deletion mapping within 16 bp of the fragment end was removed. These final filtered sets of normalized fragment level interaction frequencies were then processed using a sliding window approach with the window size of 5 kb and step size of 1 kb using the average fragment interaction frequency over the 5-kb window. These sliding interaction frequency files were then quantile normalized across all replicates in order for comparison between experiments using the “normalize.quantiles.robust” function (with use.median = TRUE) in the “preprocessCore” library in R. For display purposes,

the average of two replicates was converted to bedGraph format and displayed in the UCSC genome browser.

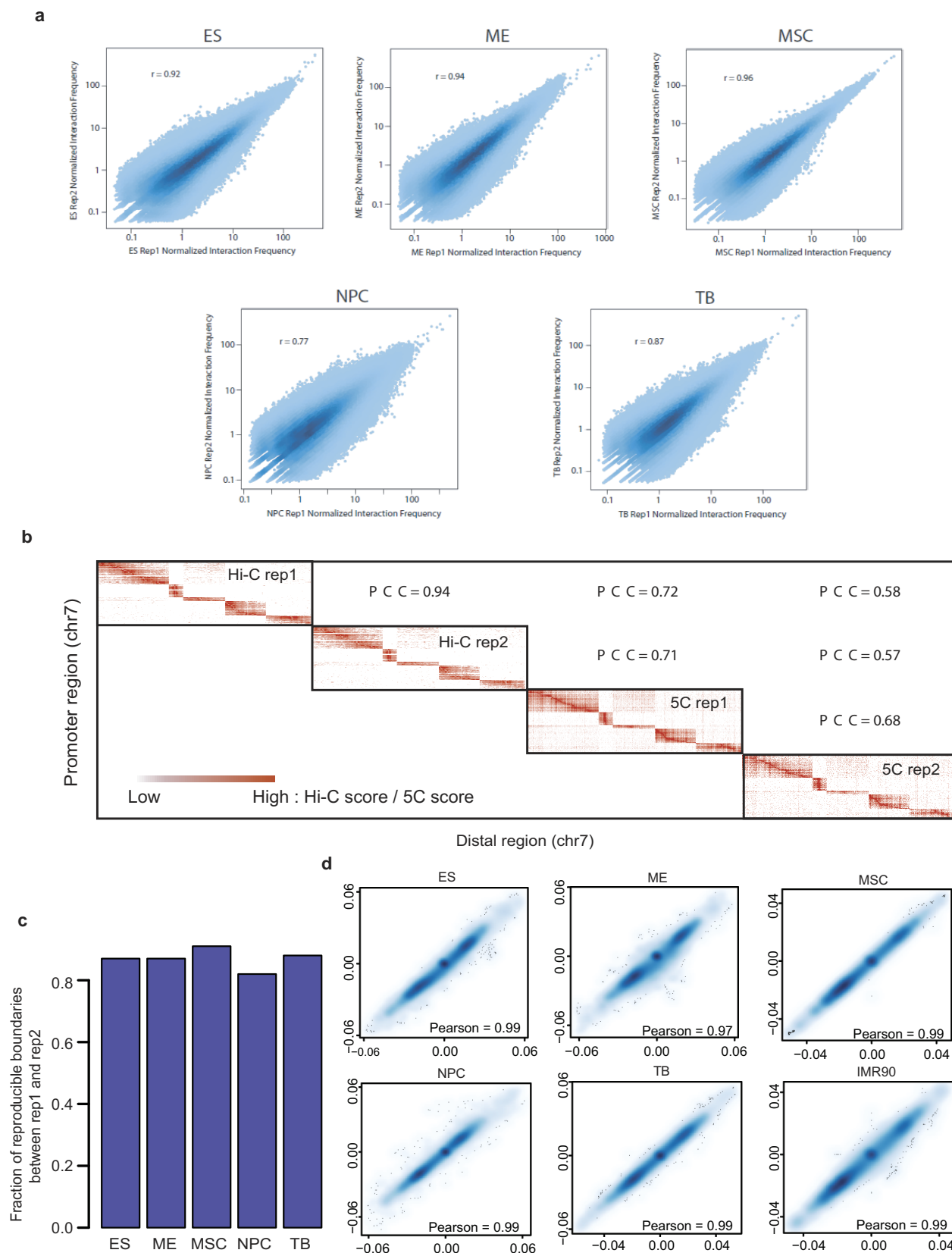
To identify regions that showed specific interactions with the bait region controlling for the genomic distance between loci, we developed a LOWESS regression model. We pooled the sliding window interaction frequency files from each of the 4C-seq replicates and performed LOWESS regression in R with the function “lowess” (with $f = 0.01$) on the \log_{10} transformed interaction frequencies controlling for the distance between the bait and potential interaction locus. We considered any region as showing ‘specific’ interactions if it showed an increase in interaction frequency greater than 2.5-fold over expected given the distance between the bait and target loci. These were considered to be the bait interacting regions.

To test for any allelic bias in 4C-seq interaction frequencies, the average normalized fragment level interaction frequency was calculated for each allele of each replicate over the bait interacting regions nearest to the transcription start site (TSS) of the putative target gene. A *t*-test was performed using these average values ($n = 2$ for each allele) to determine statistical significance.

The primers used for each 4C-seq experiment are listed in Supplementary Table 5 (please see Supplementary Table 4 for information regarding the bait regions

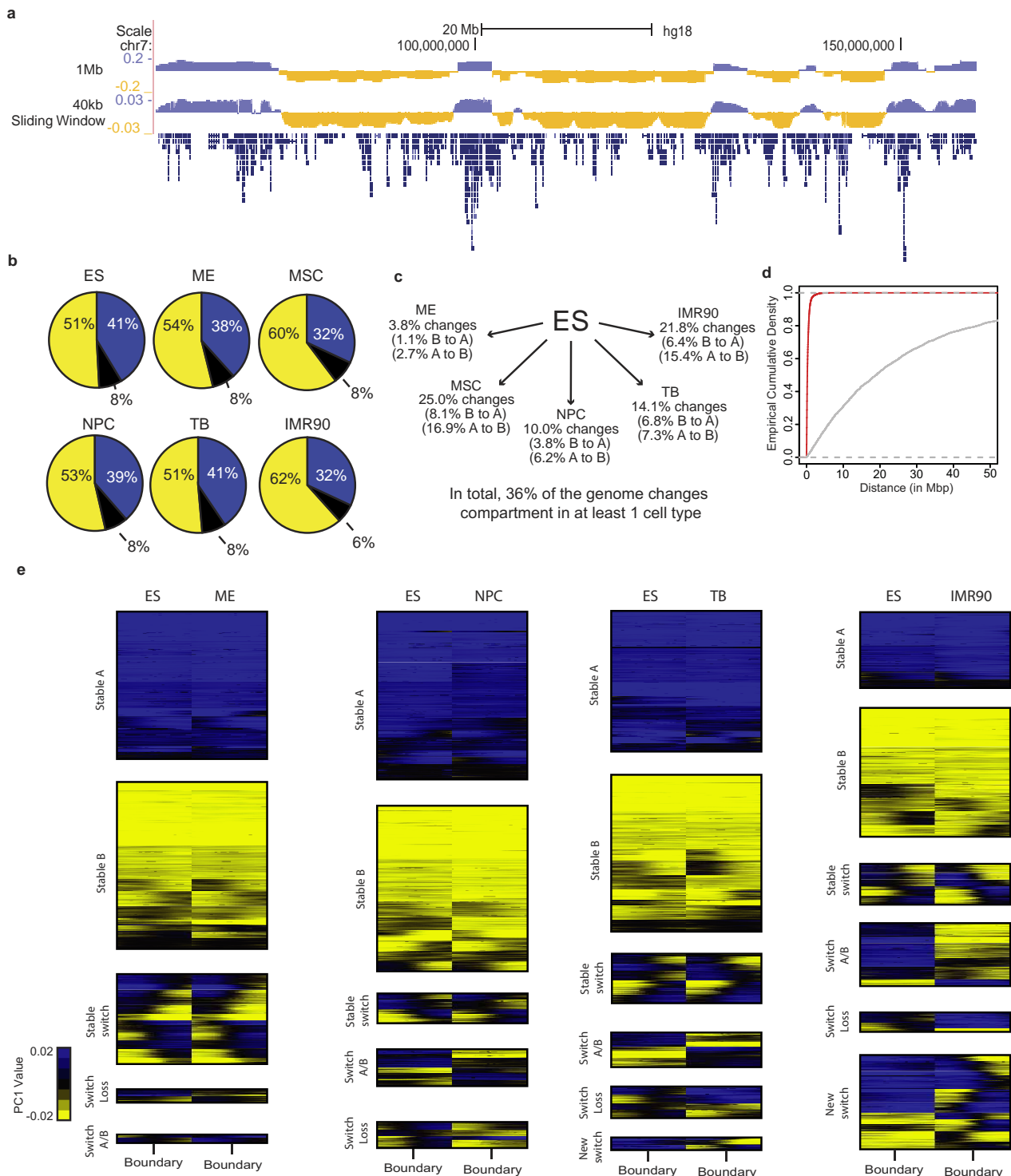
for each experiment). In Supplementary Table 5, the Illumina barcode adaptors are shown in red, with the region matching the bait region shown in blue. There is also an additional variable region in the Illumina TruSeq adaptors that has been incorporated and shown in green. The phosphorothioate bond is indicated by an asterisk.

40. Kim, T. H. *et al.* Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell* **128**, 1231–1245 (2007).
41. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genet.* **43**, 491–498 (2011).
42. Bansal, V. & Bafna, V. HapCUT: an efficient and accurate algorithm for the haplotype assembly problem. *Bioinformatics* **24**, i153–i159 (2008).
43. Browning, B. L. & Browning, S. R. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* **84**, 210–223 (2009).
44. Rajagopal, N. *et al.* RFECs: a random-forest based algorithm for enhancer identification from chromatin state. *PLOS Comput. Biol.* **9**, e1002968 (2013).
45. Degner, J. F. *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**, 390–394 (2012).



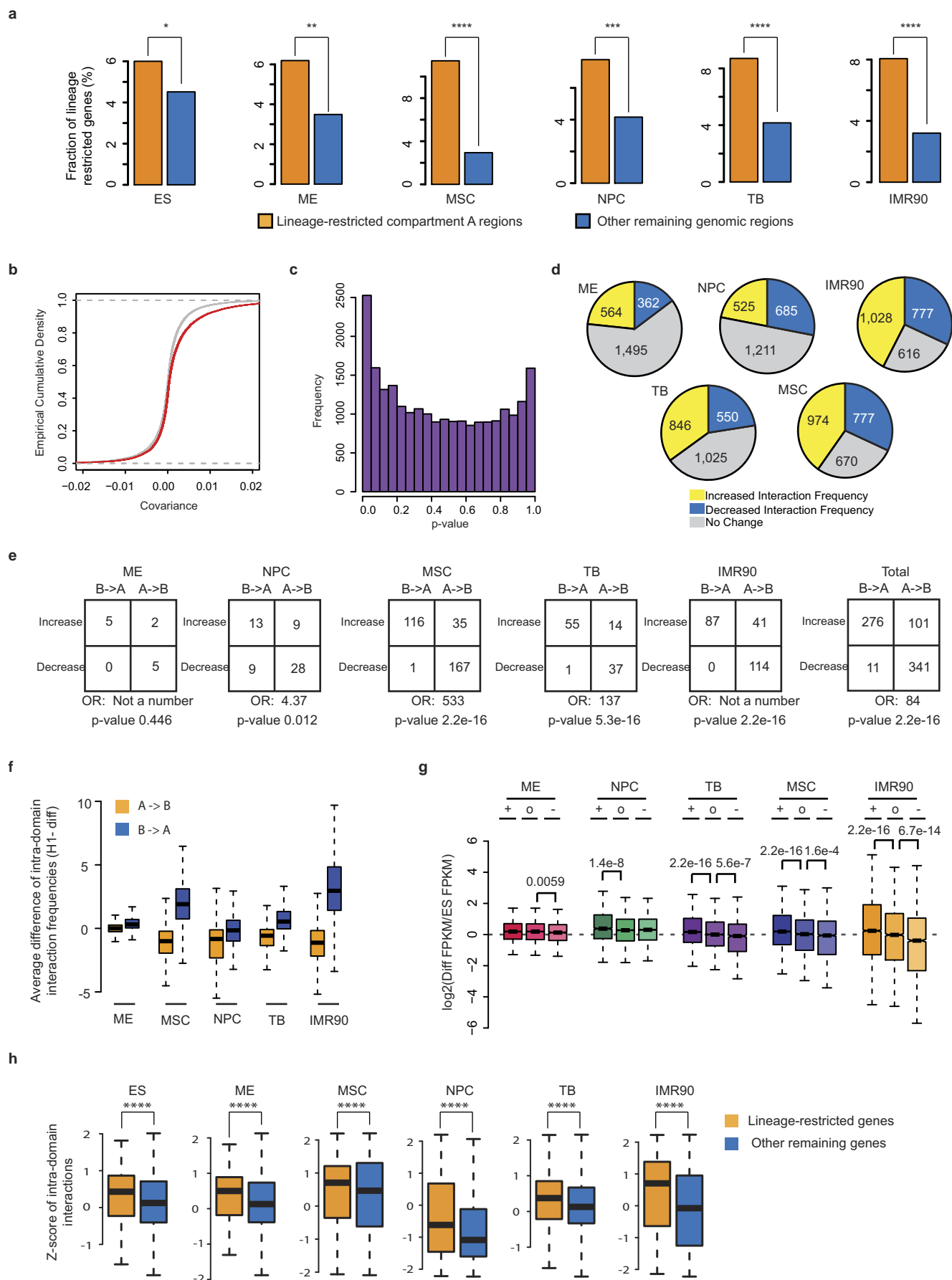
Extended Data Figure 1 | Reproducibility of Hi-C data. **a**, Scatter plots showing the correlation of Hi-C interaction frequencies between two biological replicates for H1 ES cells and H1-derived lineages. The Pearson correlation coefficient between replicates is shown in each plot. **b**, Heat maps showing interaction frequencies of Hi-C and 5C data over the chromosome 7 ENCODE loci. Pearson correlation coefficients between heat maps are shown together. The correlation coefficients between Hi-C data and 5C data (PCC 0.72, 0.71,

0.58, 0.57) are similar to the correlation coefficient between two biological replicates of 5C data (PCC 0.68). **c**, Bar plots showing the fraction of topological domain boundaries that are reproducible between biological replicates over H1 and H1-derived lineages (x axis). **d**, Scatter plot showing the PC1 values derived from compartment A/B analysis between biological replicates. PC1 values are used to determine the A and B compartments in each cell type. The Pearson correlation coefficient is shown in each graph.



Extended Data Figure 2 | A/B compartments changes are concordant with topological domain boundaries. **a**, Genome browser image of the A/B compartments determined using the previously described 1-Mb bin algorithm (1-Mb track) compared with the 40-kb sliding window approach used in our analysis (40-kb sliding window track). **b**, Pie-charts demonstrating the fraction of the genome in the A (blue) or B (yellow) compartments in each of the six lineages studied. Shown in black are regions with a PC1 of zero, often corresponding to centromeric and telomeric regions of the chromosomes. **c**, Percentage of the genome that changes A/B compartment upon differentiation of ES cells into each of the five differentiated lineages. **d**, Cumulative density plot of the distance between topological domain boundaries and transition points between the A and B compartments. The red line represents the observed distances and the grey line represents distances for

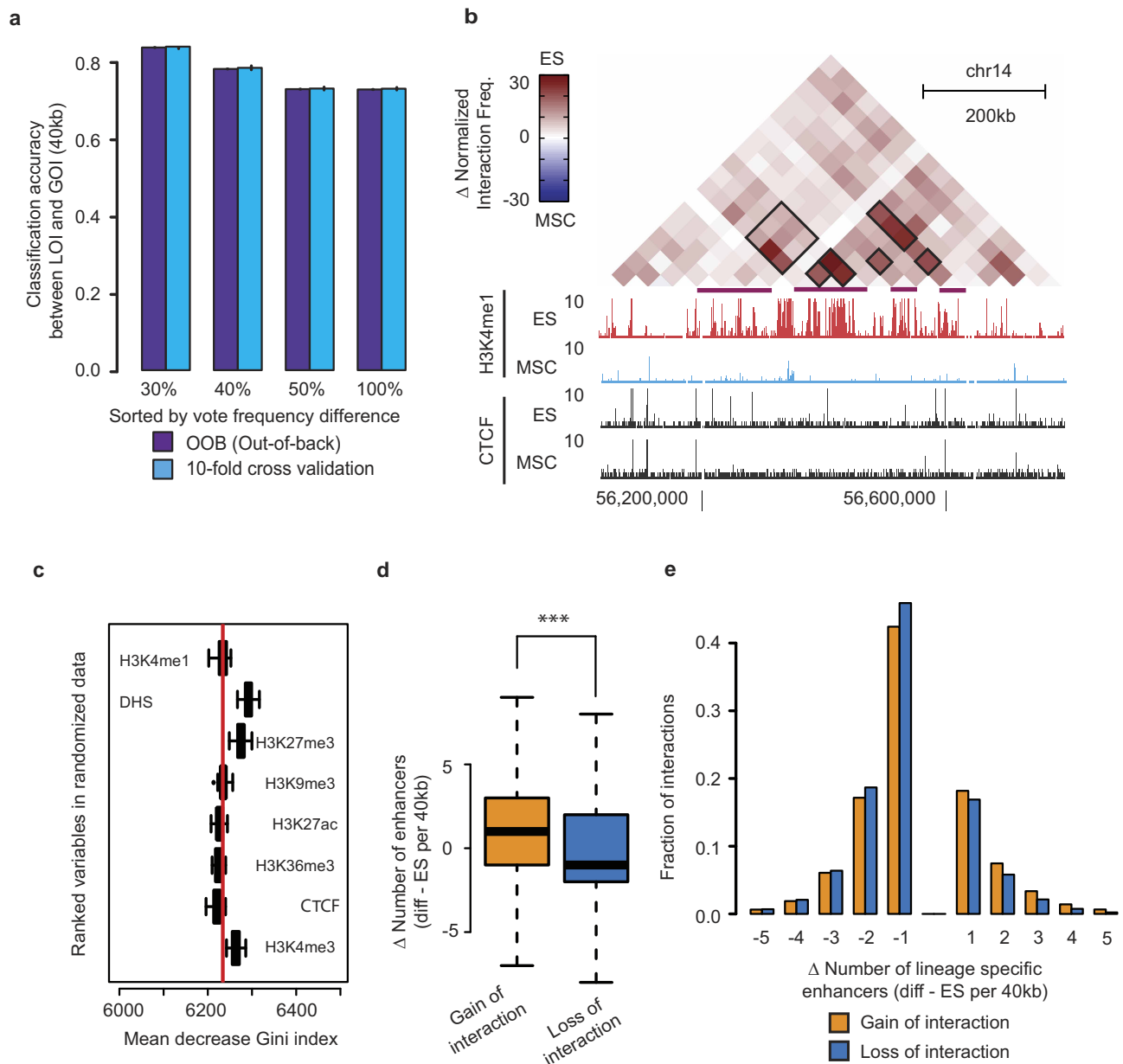
randomly generated topological domain boundaries. Domain boundaries are closer to A/B compartment transitions when compared with random (P value 2.2×10^{-16} , Wilcoxon rank sum test). **e**, K -means clustering of PC1 values in human ES cells and differentiated lineages surrounding topological domain boundaries. Similar to Fig. 1c, domain boundaries correspond to the transition points between the A/B compartments, and changes in A/B compartments that occur during differentiation tend to occur at domain boundaries. Regions that stay as A or B compartment are termed stable A or stable B. Regions that stay as A/B compartment switching are labelled as stable switch. Regions where the boundary becomes a new switching point for the A/B compartment are labelled new switch. Regions that previously were A/B compartment switching but are no longer after differentiation are labelled switch loss. Regions that entirely switch from A to B or vice versa are labelled as switch A/B.



Extended Data Figure 3 | A/B compartment changes and gene expression.

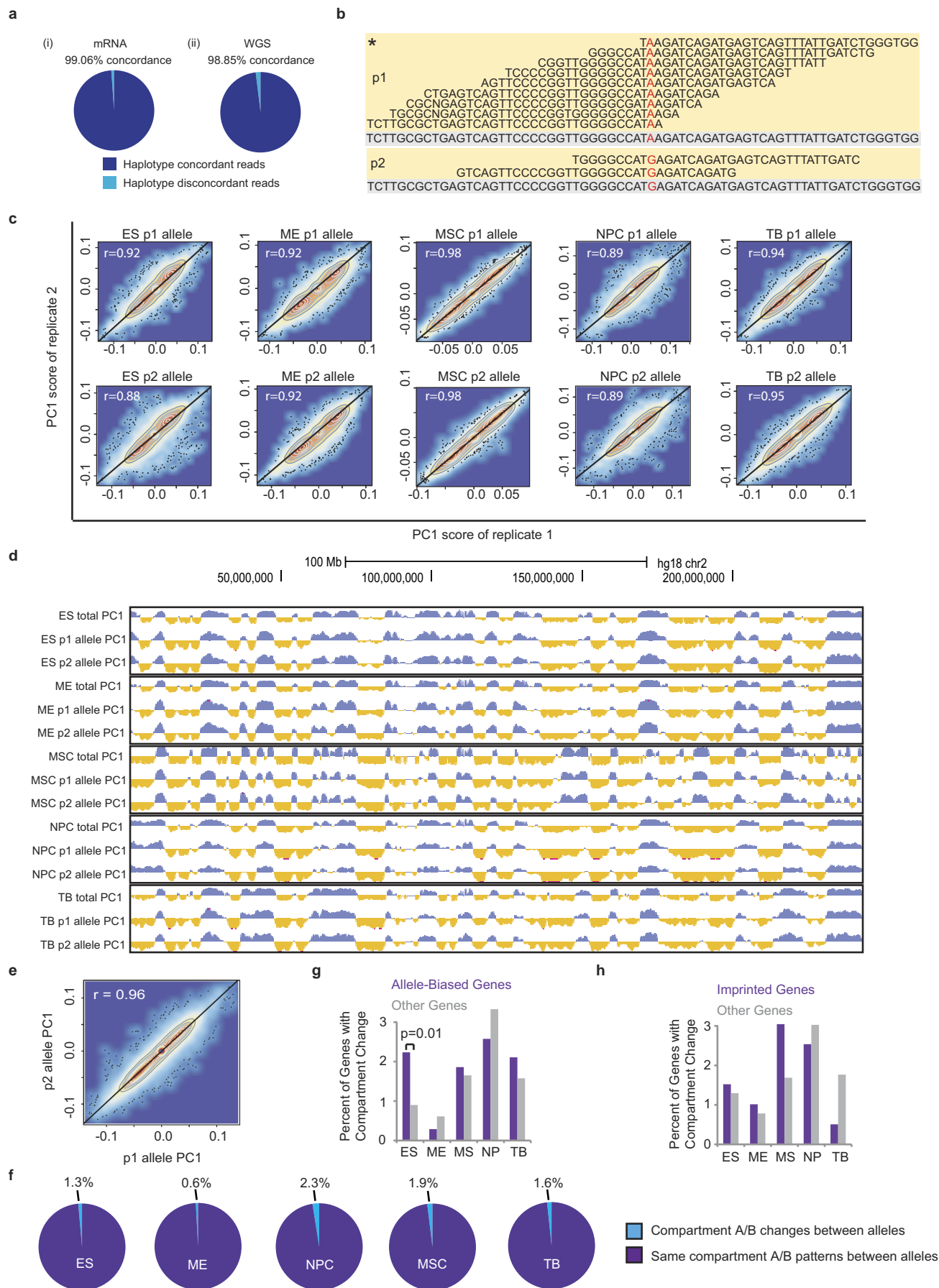
a, Fraction of lineage-restricted genes in lineage-restricted compartment A regions and other remaining regions. If only one or two cell lines are assigned as compartment A across the six lineages, the region is defined as a lineage-restricted compartment A region. For all six lineages, lineage-restricted genes tend to be enriched in lineage-restricted compartment A regions compared to other genomic regions (P values < 0.05 , Fisher's exact test). **b**, Empirical cumulative density plot of covariance values between gene expression and PC1 score across the 6 lineages analysed. Shown in red are the observed covariance values, while in grey are covariance values calculated after randomly shuffling the vector of gene expression values for each lineage. The slight shift of the red curve to the right indicates that the observed data has a subset of genes with higher covariance values than would be expected at random, indicating that a subset of genes have concordant gene expression and PC1 values. **c**, Histogram of P values for the covariance between gene expression and PC1 values for each gene. To calculate the P value, a random background distribution of covariance values was generated by calculating the covariance between the PC1 values and a randomly shuffling of the vector of gene expression values for each gene. This shuffling was performed 1,000 times. The actual observed covariance can then be assigned a rank based P value given the random background distribution for that gene. The plot shows that a subset of genes is enriched for having low P values, consistent with the idea that a subset of genes shows concordant gene expression and compartment status. **d**, Pie charts showing the fraction of domains that are identified as having a concerted increase (yellow) or decrease (blue) in intra-domain interaction frequency between H1 human ES cells and the five lineages analysed.

e, Relationship between A/B compartment and intra-domain interactions. 2×2 tables for each lineage (or for all lineages combined, labelled as total) for domains that show a concerted increase or decrease in interaction frequency and whether they show a change from A to B or B to A compartments. Domains are considered to undergo a compartment change if $> 80\%$ of individual bins within the domain change compartment. Odds ratio (OR) for each lineage and the total are listed, as are P values for the association (Fisher's exact test). **f**, Box plots showing the average difference of intra-domain interaction frequencies between H1 human ES cells and the five lineages analysed. Regions that change from compartment B to A (blue) tend to show increased intra-domain interaction frequencies compared to regions that change from compartment A to B (orange). P values are less than 2.2×10^{-16} for all lineages (KS test). **g**, Box plots of the fold-change in gene expression of genes located in domains that have a significant increase (+), decrease (−), or no change (0) in intra-domain interaction frequency between ES cells and each of the differentiated lineages specified. The fold-change in expression is the \log_2 of the expression of a gene in the differentiated cells over ES cells (P values from Wilcoxon rank-sum test). **h**, Box plots showing Z-scores of intra-domain interactions between lineage-restricted genes and other remaining genes. The average intra-domain interaction frequency was calculated for each domain in six lineages analysed and converted to a Z-score. The Z-score of each gene was assigned by the Z-score of corresponding domain that includes the gene. Lineage-restricted genes tend to reside in domains with higher Z-scores compared to other remaining genes. The P values were less than 1×10^{-4} from the KS test.



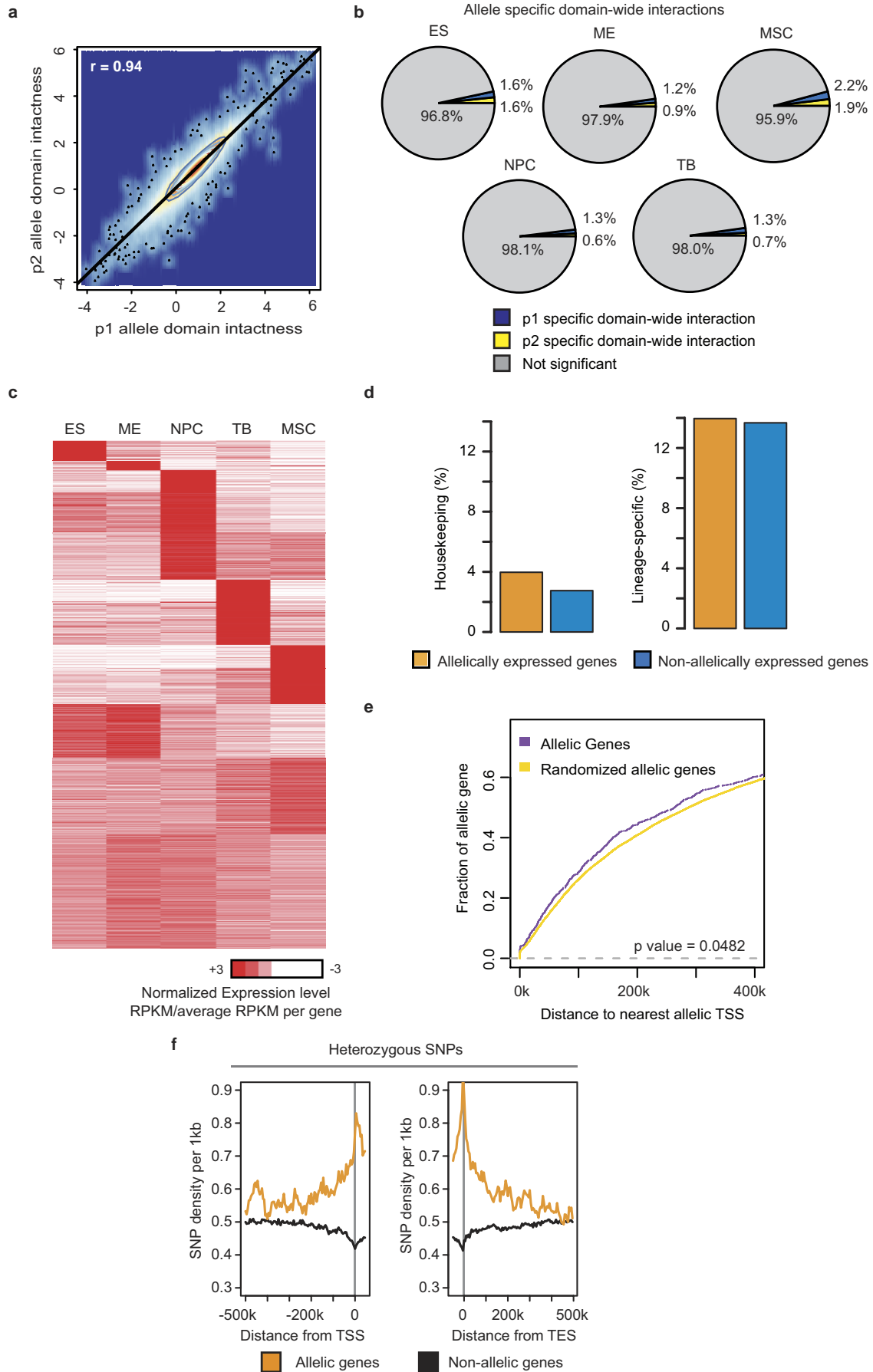
Extended Data Figure 4 | Random Forest model to predict Hi-C interaction changes. **a**, Comparison of the classification accuracy between tenfold cross validation and the out-of-back (OOB) error rates. The two methods show similar classification accuracies at each vote frequency threshold. **b**, Heat map showing the difference of normalized interaction frequencies between H1 and MS cells. The boxes indicate the regions with relatively strong higher interaction frequencies in H1 ES cells. H3K4me1 and CTCF ChIP-seq signals are also shown together. H3K4me1 ChIP-seq signals are highly correlated with changes of interaction frequency. **c**, Similar to Fig. 2e, ranked Gini index of different chromatin features of the Random Forest model when using randomized data. The red line represents the centre of the Gini index.

d, Box plots demonstrating the difference in the number of enhancers in each 40-kb bin that undergoes a gain of interactions (GOI) or loss of interactions (LOI) upon differentiation. We observe that regions that are involved gain of interactions tend to contain more enhancers in differentiated lineages compared to H1 cells. We only considered regions containing more than 10 enhancers in total for H1 and differentiated cell lines (** P value $< 2.2 \times 10^{-16}$, KS test). **e**, Histogram showing the fraction of interactions classified as GOI (orange) or LOI (blue) according to the difference of the number of lineage-specific enhancers between differentiated lineages and H1 ES cells. The regions with more lineage-specific enhancers in differentiated lineages are enriched by gain of interactions.



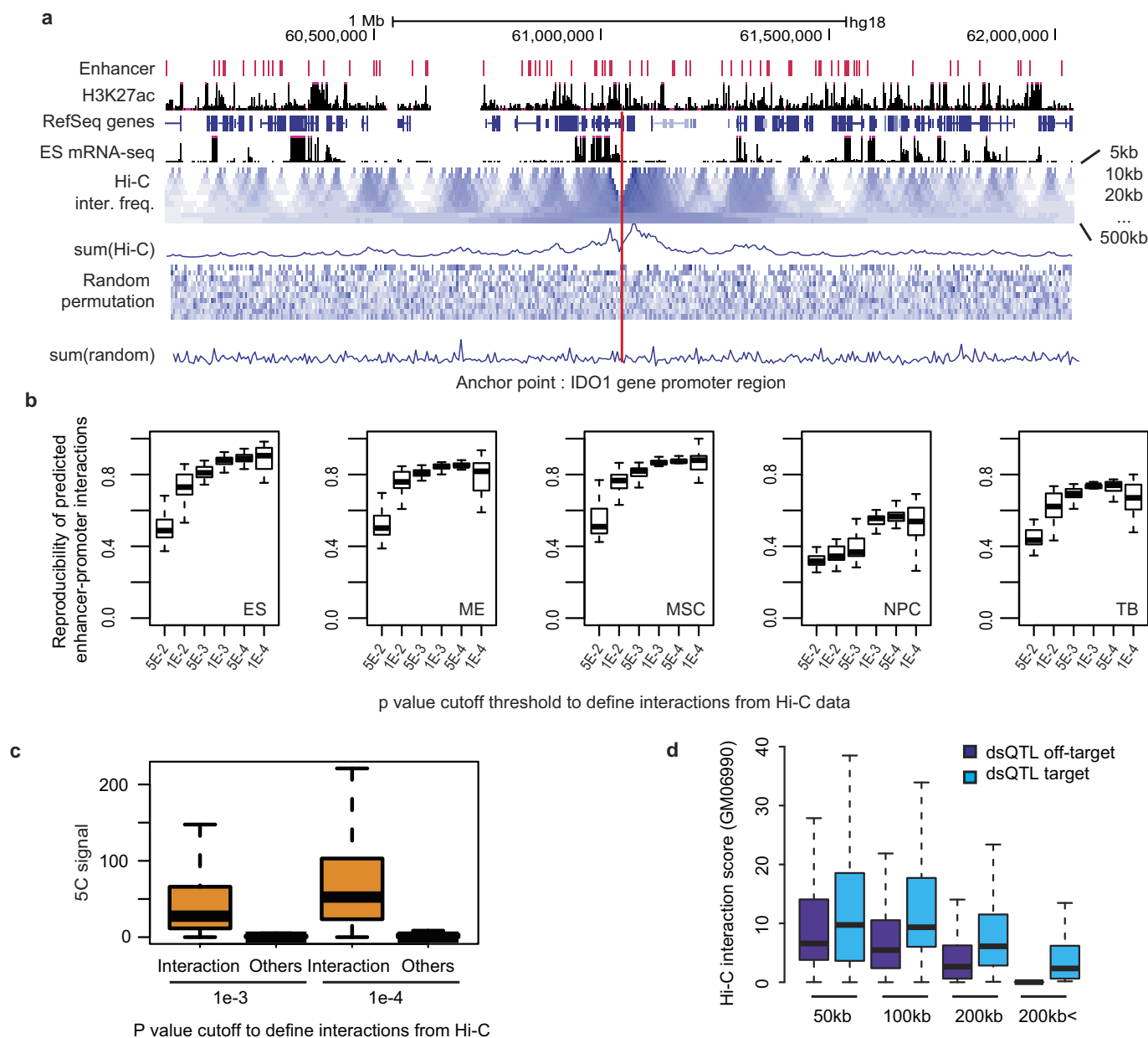
Extended Data Figure 5 | Allele-specific chromatin structure. **a**, Validation of haplotypes by (i) RNA-sequencing (i) and whole-genome sequencing (WGS) (ii). Shown in dark blue is the percentage of reads connecting variants in the same predicted haplotype, while in light blue is the percentage of reads connecting variants predicted to be on different haplotypes. **b**, Inset labelled with an asterisk is from Fig. 3b showing DHS sequencing reads over a SNP upstream from the *SNRPN* gene, demonstrating how different chromatin features can be assigned to a given haplotype. **c**, Scatter plots showing the correlation coefficient of PC1 values obtained from compartment A/B analysis between the two biological replicates for each allele. Despite the reduction in reads when Hi-C data are split into two alleles, the PC1 scores were highly reproducible between replicates. **d**, Shown is a genome browser image of PC1 values in chromosome 2 for the p1 allele, p2 allele, and for all Hi-C reads

without resolving the two alleles. PC1 scores are highly consistent, suggesting that homologous chromosomes fold in highly similar patterns. **e**, Scatter plot of PC1 values between the p1 and p2 alleles in H1 and H1-derived lineages. The Pearson correlation coefficient value is 0.96. **f**, Fraction of the genome that shows changes in A/B compartment status across alleles. For ES, ME, MS, NP and TB cells, 1.3%, 0.6%, 1.9%, 2.3% and 1.6% of total genomic regions shows allelic compartment A/B patterns, respectively. **g**, Percentage of allele-biased (purple) or non-allele-biased (grey) genes that have different A/B compartment status in each lineage. Only in ES cells is there a significant association between allele-biased genes and regions with variable A/B compartment between alleles (Fisher's exact test). **h**, Similar to **g**, but showing the association between imprinted genes and changes in A/B compartment between alleles. No lineage shows a significant association.



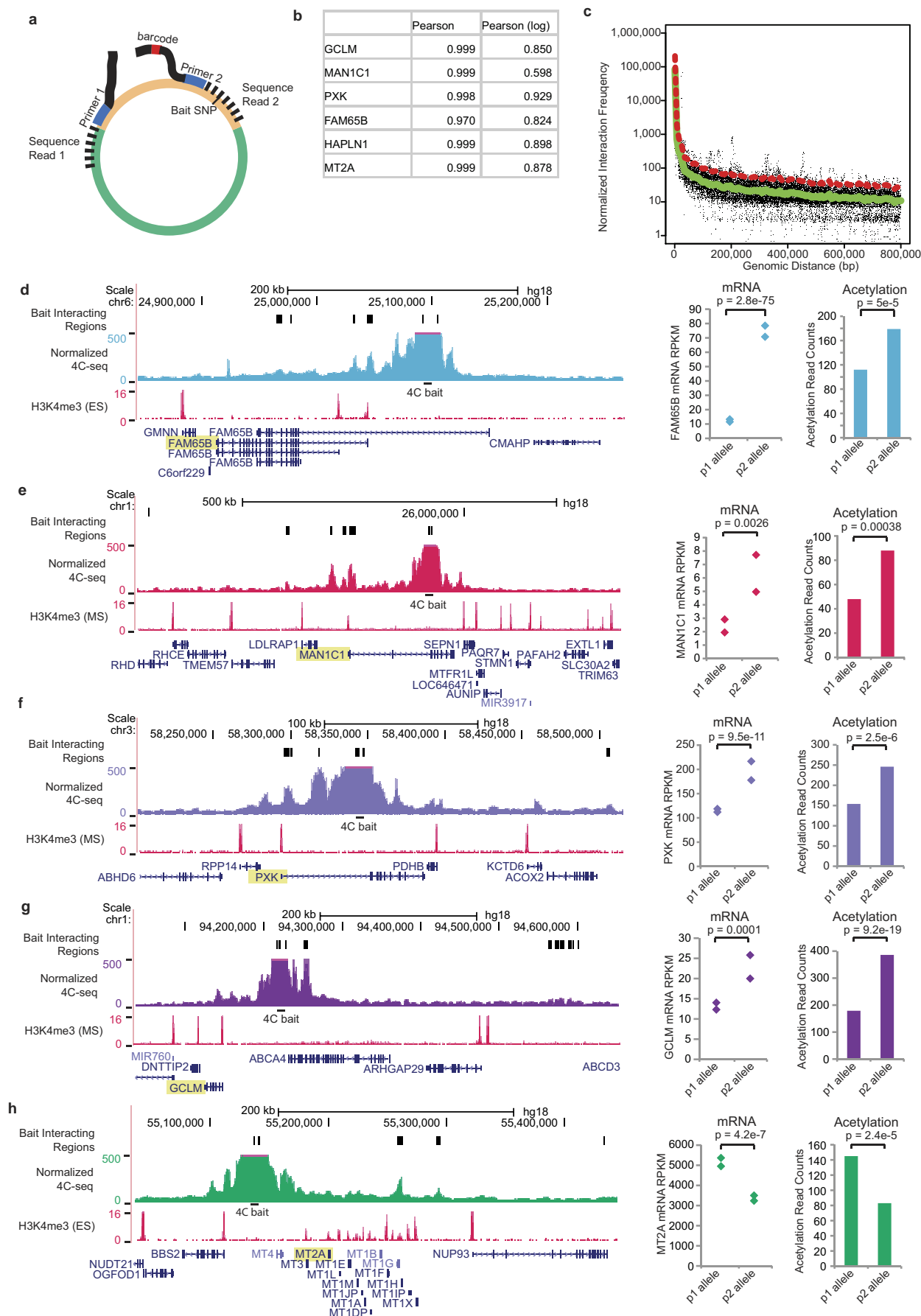
Extended Data Figure 6 | Domain-wide structural changes and allele-biased genes. **a**, Scatterplot showing domain ‘intactness’ between the p1 and p2 alleles. Domain intactness is defined as the \log_2 ratio of the total number of intra-domain interactions versus total number of inter-domain interactions for each topological domain. The highly correlated domain intactness scores between the p1 and p2 alleles support the similar topological domain patterns between two homologous chromosomes. **b**, Pie charts showing the fraction of domains that are identified as having a concerted p1 allele specific increase (blue) or p2 allele specific increase (yellow) in interaction frequency. Grey in the pie charts indicates the fraction of domains that do not show allele specific patterns compared to the random model (P value cutoff is 0.001). **c**, Heat map showing K -means clustering ($k = 12$) of gene expression levels of allele-biased genes across each of the five H1 lineages. The expression levels are shown as the fold-change of expression in each lineage relative to the average expression

level across each of the five lineages. Allele-biased genes consist of both cell-type specific and constitutively expressed genes. **d**, Fraction of housekeeping genes and lineage-restricted genes that show allele-biased expression. There is no statistically significant enrichment between allele-biased genes (orange) and non-allele-biased genes (blue) among housekeeping or lineage-restricted genes. **e**, Empirical cumulative density plot of the distance between each allele-biased gene and the nearest allele-biased gene (purple) as compared with randomly chosen genes (yellow). The difference from an allele-biased gene to the nearest allele-biased gene is less than what would be expected at random ($P = 0.0482$, Wilcoxon rank sum test), however, the difference is subtle, indicating that most allele-biased expression does not occur in clusters. **f**, Rate of heterozygous SNPs near both allele-biased (gold) and non-allele-biased (black) genes. See Supplementary Information for further details.



Extended Data Figure 7 | Identification of enhancer–promoter interactions. **a**, Shown is the Hi-C interaction frequency between the *IDO1* gene promoter regions and ± 1 Mb surrounding regions. Each entry in the heat map of Hi-C inter. freq. indicates Hi-C interaction frequency between the promoter and the surrounding regions. Each row indicates the Hi-C interaction frequencies for a given window size. The heat map of random permutation was generated by randomizing each row in Hi-C interaction frequency. The sum (Hi-C) and sum (random) indicate summation of Hi-C interaction frequencies for each 5-kb window. Predicted enhancers, H3K27ac, RNA-seq, and RefSeq gene information are shown together. **b**, Box plots showing the reproducibility of predicted enhancer–promoter interactions between two

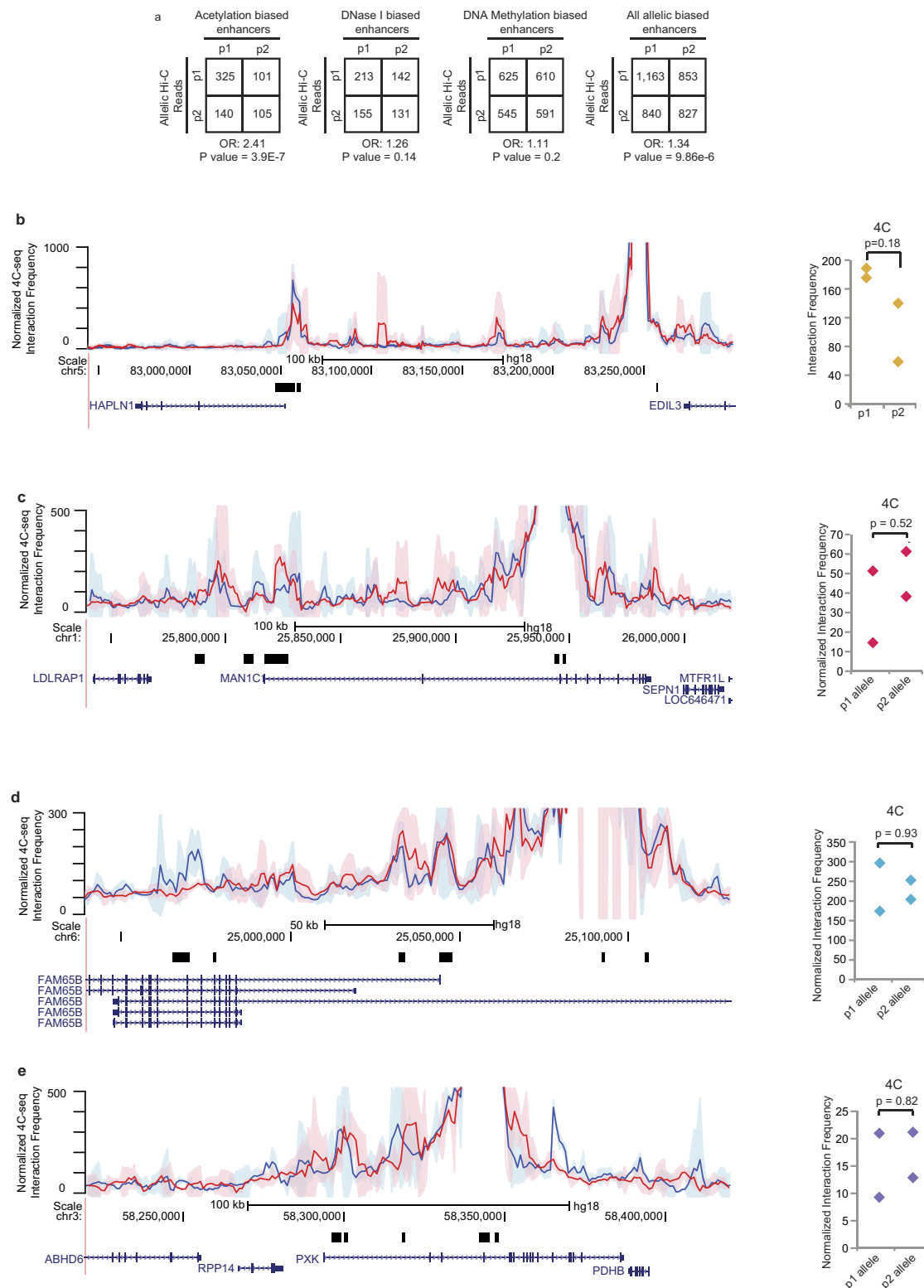
biological replicates for each chromosome with different P value cutoffs. If the P value is less than 0.001, the reproducibility between replicates is over 80%. **c**, Distribution of 5C signals between interacting pairs (interaction) and non-interacting pairs (others) defined by Hi-C interaction frequency score with different P value cutoffs. Interaction pairs defined by Hi-C interactions are also strongly enriched by 5C signals at both P value cutoffs ($n = 11,461$ for 1×10^{-3} and $n = 1,841$ for 1×10^{-4}). **d**, Relationship between Hi-C interaction frequency scores and dsQTL target-gene pairs according to distance between gene and its target DHS regions. Target-gene relationships tend to show higher Hi-C interaction frequency scores compared to off-target-gene relationships.



Extended Data Figure 8 | 4C-seq between allelic enhancers and allelic genes.

a, Diagram of experimental design for 4C-seq and allelic 4C-seq. The orange region depicts the 4C bait locus, and the green region is the interacting target region. Primers containing the Illumina adaptor sequences and a bait-specific sequence are used for inverse PCR of the target region. Barcodes based on the Illumina TruSeq adaptors are incorporated into the primer sequences to allow for multiplexing. The second primer will read a sequence from the bait region with a SNP that determines the allele from which the bait was derived. **b**, Pearson correlation coefficients between replicates for each of the loci tested. Also shown is the Pearson correlation coefficient between replicates after log-transformation of the interaction frequency. **c**, Scatter plot of LOWESS regression of 4C-seq data. The x axis shows the genomic distance between the bait region and the putative target region. The y axis is the \log_{10} of the quantile normalized interaction frequencies. LOWESS was performed to generate an expected interaction frequency at each genomic distance (green line). A cut off of 2.5-fold over expected (shown in the red dashed line) is used to determine if a region shows specific interactions, so-called bait interacting

regions (BIRs). **d**, Normalized 4C-seq interaction frequencies surrounding a bait region located in an allelic enhancer near the *FAM65B* gene. The location of the bait is labelled as 4C bait. Regions with significant interactions according to the LOWESS regression model are labelled as black lines in the track marked bait interacting regions. Shown to the right is the level of mRNA-seq data for each allele of the *FAM65B* gene, the level of histone acetylation at the allelic enhancer bait region. Significance for mRNA-seq data was calculated using the edgeR software package in R. Acetylation P values were calculated using a binomial test. **e**, Similar to **d**, but for a 4C seq bait located in the *MAN1C1* gene. **f**, Similar to **d**, but for an allelic enhancer located in the *PXK* gene. **g**, Similar to **d**, but for an enhancer located in near the *GCLM* gene. Of note, this allele-biased enhancer forms no specific contacts with any allelic genes. **h**, Similar to **d**, but for an enhancer located near the *MT2A* gene. There are no specific interactions between the allelic enhancer and the *MT2A* gene. There are specific interactions between the allelic enhancer and the *MT1H* and *MT1G* genes. However, neither gene has an exonic SNP and therefore we cannot determine if these genes have allele-biased expression.



Extended Data Figure 9 | 4C-seq interacting regions from allelic enhancers.

a, Allelic Hi-C interaction reads shown for allelic gene-enhancer pairs defined using either allelic histone acetylation, DHS or DNA methylation. Odds ratios (OR) and *P* values (Fisher's exact test) are shown. For enhancers defined by histone acetylation and the pooled set of enhancers, a statistically significant association between allele-biased Hi-C reads and allele-biased enhancer activity is observed. **b**, Normalized 4C-seq interaction frequencies surrounding a bait region located in an allelic enhancer near the *HAPLN1* gene. The blue line shows the interaction frequency for the p1 allele and the red line shows interaction frequencies for the p2 allele. The shaded regions represent 95%

confidence intervals for the interaction frequency. Shown to the right are the allele-specific normalized 4C interaction frequencies for each allele. 4C-seq interaction frequencies for each allele were computed over the significant bait interacting regions nearest to the target gene TSS. Significance testing for allelic 4C-seq data was performed by *t*-test ($n = 2$ for each allele). Black bars below the plot indicate regions identified as bait-interacting regions (BIRs). Of note, the panel to the right is the same as that found in Fig. 5f. **c**, Similar to **b**, but for an allelic enhancer located in the *MAN1C* gene. **d**, Similar to **b**, but for a 4C seq bait located in the *FAM65B* gene. **e**, Similar to **b**, but for an enhancer located in near the *PDK* gene.

Genetic and epigenetic fine mapping of causal autoimmune disease variants

Kyle Kai-How Farh^{1,2*}, Alexander Marson^{3*}, Jiang Zhu^{1,4,5,6}, Markus Kleinewietfeld^{1,7†}, William J. Housley⁷, Samantha Beik¹, Noam Shores¹, Holly Whitton¹, Russell J. H. Ryan^{1,5}, Alexander A. Shishkin^{1,8}, Meital Hatan¹, Marlene J. Carrasco-Alfonso⁹, Dita Mayer⁹, C. John Luckey⁹, Nikolaos A. Patsopoulos^{1,10,11}, Philip L. De Jager^{1,10,11}, Vijay K. Kuchroo¹², Charles B. Epstein¹, Mark J. Daly^{1,2}, David A. Hafler^{1,7§} & Bradley E. Bernstein^{1,4,5,6§}

Genome-wide association studies have identified loci underlying human diseases, but the causal nucleotide changes and mechanisms remain largely unknown. Here we developed a fine-mapping algorithm to identify candidate causal variants for 21 autoimmune diseases from genotyping data. We integrated these predictions with transcription and *cis*-regulatory element annotations, derived by mapping RNA and chromatin in primary immune cells, including resting and stimulated CD4⁺ T-cell subsets, regulatory T cells, CD8⁺ T cells, B cells, and monocytes. We find that ~90% of causal variants are non-coding, with ~60% mapping to immune-cell enhancers, many of which gain histone acetylation and transcribe enhancer-associated RNA upon immune stimulation. Causal variants tend to occur near binding sites for master regulators of immune differentiation and stimulus-dependent gene activation, but only 10–20% directly alter recognizable transcription factor binding motifs. Rather, most non-coding risk variants, including those that alter gene expression, affect non-canonical sequence determinants not well-explained by current gene regulatory models.

Genome-wide association studies (GWAS) have revolutionized the study of complex human traits by identifying thousands of genetic loci that contribute susceptibility for a diverse set of diseases^{1,2}.

However, progress towards understanding disease mechanisms has been limited by difficulty in assigning molecular function to the vast majority of GWAS hits that do not affect protein-coding sequence. Efforts to decipher biological consequences of non-coding variation face two major challenges. First, due to haplotype structure, GWAS tend to nominate large clusters of single nucleotide polymorphisms (SNPs) in linkage disequilibrium (LD), making it difficult to distinguish causal SNPs from neutral variants in linkage. Second, even assuming the causal variant can be identified, interpretation is limited by incomplete knowledge of non-coding regulatory elements, their mechanisms of action, and the cellular states and processes in which they function.

Inflammatory autoimmune diseases, which reflect complex interactions between genetic variation and environment, are important systems for genetic investigation of human disease³. They share a substantial degree of immunopathology, with increased activity of auto-reactive CD4⁺ T cells secreting inflammatory cytokines and loss of regulatory T-cell (T_{reg}) function⁴. A critical role for B cells in certain diseases has also been revealed with the therapeutic efficacy of anti-CD20 antibodies⁵. Immune homeostasis depends on a balance of CD4⁺ pro-inflammatory (Th1, Th2, Th17) cells and FOXP3⁺ suppressive T_{regs}, each of which expresses distinct cytokines and surface molecules⁶. Each cell type is controlled by a unique set of master transcription factors (TFs) that



EPIGENOME ROADMAP

A *Nature* special issue
nature.com/epigenomeroadmap

directly shape cell-type-specific gene expression programs, which include genes implicated in autoimmune diseases^{7–9}. Immune subsets also have characteristic *cis*-regulatory landscapes, including distinct sets of enhancers that may be distin-

guished by their chromatin states^{9–13} and associated enhancer RNAs (eRNA)¹⁴. Familial clustering of different autoimmune diseases suggests that heritable factors underlie common disease pathways, although disparate clinical presentations and paradoxical effects of drugs in different diseases support key distinctions¹⁵.

GWAS have identified hundreds of risk loci for autoimmunity¹⁵. Although most risk variants have subtle effects on disease susceptibility, they provide unbiased support for possible aetiological pathways, including antigen presentation, cytokine signalling, and NF-κB transcriptional regulation¹⁵. The associated loci are enriched for immune cell-specific enhancers^{10,16,17} and expansive enhancer clusters^{18,19}, termed ‘super-enhancers’, implicating gene regulatory processes in disease aetiology. However, as is typical of GWAS, the implicated loci comprise multiple variants in LD and rarely alter protein-coding sequence, which complicates their interpretation.

Here, we integrated genetic and epigenetic fine mapping to identify causal variants in autoimmune disease-associated loci and explore their functions. Based on dense genotyping data²⁰, we developed a novel algorithm to predict for each individual variant associated with 21 autoimmune diseases, the likelihood that it represents a causal variant. In parallel, we generated *cis*-regulatory element maps for a spectrum of immune cell types. Remarkably, ~60% of likely causal variants map to

¹Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA. ²Analytical and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. ³Diabetes Center and Division of Infectious Diseases, Department of Medicine, University of California, San Francisco, California 94143, USA. ⁴Howard Hughes Medical Institute, Chevy Chase, Maryland 20815, USA.

⁵Department of Pathology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts 02114, USA. ⁶Center for Systems Biology and Center for Cancer Research, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. ⁷Departments of Neurology and Immunobiology, Yale School of Medicine, New Haven, Connecticut 06511, USA. ⁸California Institute of Technology, 1200 E California Boulevard, Pasadena, California 91125, USA. ⁹Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA. ¹⁰Program in Translational NeuroPsychiatric Genomics, Institute for the Neurosciences, Department of Neurology, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts 02142, USA.

¹¹Division of Genetics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts 02142, USA. ¹²Center for Neurologic Diseases, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts 02142, USA. [†]Present address: Translational Immunology, Medical Faculty Carl Gustav Carus, TU Dresden, 01307 Dresden, Germany.

*These authors contributed equally to this work.

§These authors jointly supervised this work.

enhancer-like elements, with preferential correspondence to stimulus-dependent CD4⁺ T-cell enhancers that respond to immune activation by increasing histone acetylation and transcribing non-coding RNAs. Although these enhancers frequently reside within extended clusters, their distinct regulatory patterns and phenotypic associations suggest they represent independent functional units. Causal SNPs are enriched near binding sites for immune-related transcription factors, but rarely alter their cognate motifs. Our study provides a unique resource for the study of autoimmunity, links causal disease variants with high probability to context-specific immune enhancers, and suggests that most non-coding causal variants act by altering non-canonical regulatory sequence rather than recognizable consensus transcription factor motifs.

Fine-mapped genetic architecture of disease

To explore the genetic architecture underlying common diseases, we collected 39 well-powered GWAS studies (Methods). Clustering of diseases and traits based on their shared genetic loci revealed groups of phenotypes with related clinical features (Fig. 1a). This highlighted a large cluster of immune-mediated diseases forming a complex network of shared genetic loci; on average, 69% of the associated loci for each disease were shared with other autoimmune diseases, although no two diseases shared more than 38% of their loci.

We focused subsequent analysis on autoimmune diseases, reasoning that recent dense genotyping data combined with emerging approaches for profiling epigenomes of specialized immune cells would provide an opportunity to identify and characterize the specific causal SNPs. Prior studies that have integrated GWAS with epigenomic features focused on lead SNPs or multiple associated SNPs within a locus, of which only a small minority reflects causal variants^{10,16–19,21}. Although these studies demonstrated enrichments within enhancer-like regulatory elements, they could not with any degree of certainty pinpoint the specific elements or processes affected by the causal variants. To overcome this limitation, we leveraged dense genotyping data to refine a statistical model for predicting causal SNPs from genetic data alone. Rare recombination events within haplotypes can provide information on the identity of the causal SNP, provided sufficient genotyping density and sample size. We therefore

examined a cohort of 14,277 cases with multiple sclerosis and 23,605 healthy controls genotyped using the Immunochip, which comprehensively covers 1000 Genomes Project SNPs²² within 186 loci associated with autoimmunity²⁰. We developed an algorithm, Probabilistic Identification of Causal SNPs (PICS), that estimates the probability that an individual SNP is a causal variant given the haplotype structure and observed pattern of association at the locus (Methods, Extended Data Figs 1–4).

The *IFI30* locus (Fig. 1b, c) presents an illustrative example of the LD problem and the PICS strategy. The most strongly associated SNP at the locus is rs11554159 (R76Q, G>A; minor allele is protective), a missense variant in *IFI30*, which encodes a lysosomal enzyme that processes antigens for MHC presentation²³. Although dozens of SNPs at the locus are significantly associated with disease, the association for each additional SNP follows a linear relationship with its linkage to rs11554159/R76Q, suggesting they owe their association solely to linkage with this causal variant. We used permutation to estimate the posterior probability for each SNP in the locus to be the causal variant, given the observed patterns of association. Interestingly, prior GWAS studies²⁴ had attributed the signal at this locus to a missense variant in a neighbouring gene, *MPV17L2* (rs874628, $r^2 = 0.9$ to R76Q), with no known immune function. However, we find that the R76Q variant is approximately ten times more likely than rs874628 to be the causal SNP and three times more likely than the next closest SNP (a non-coding variant), providing compelling evidence that the *IFI30* missense variant is the causal variant in the locus.

We next generalized PICS to analyse 21 autoimmune diseases, using Immunochip data when they were available or imputation to the 1000 Genomes Project²² when they were not (Methods; Supplementary Table 1). We mapped 636 autoimmune GWAS signals to 4,950 candidate causal SNPs (mean probability of representing the causal variant responsible for the GWAS signal: ~10%). PICS indicates that index SNPs reported in the GWAS catalogue have on average only a 5% chance of representing a causal SNP. Rather, GWAS catalogue index SNPs are typically some distance from the PICS lead SNP (median 14 kb), and many are not in tight LD (Fig. 1d and Extended Data Fig. 5). PICS identified a single most

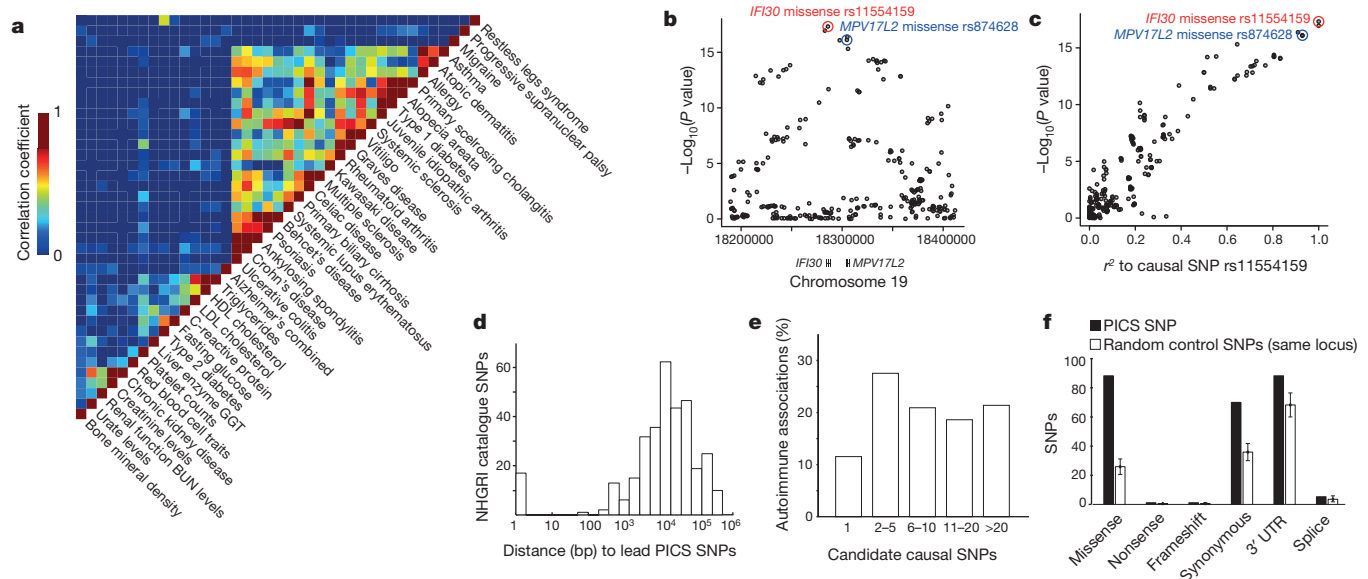


Figure 1 | Genetic fine mapping of human disease. **a**, GWAS catalogue loci were clustered to reveal shared genetic features of common human diseases and phenotypes. Colour scale indicates correlation between phenotypes (red = high, blue = low). **b**, Association signal to multiple sclerosis for SNPs at the *IFI30* locus. **c**, Scatter plot of SNPs at the *IFI30* locus demonstrates the linear relationship between LD distance (r^2) to rs11554159 (red) and association signal. **d**, Candidate causal SNPs were predicted for 21 autoimmune diseases

using PICS. Histogram indicates genomic distance (bp) between PICS Immunochip lead SNPs and GWAS catalogue index SNPs. **e**, Histogram indicates number of candidate causal SNPs per GWAS signal needed to account for 75% of the total PICS probability for that locus. **f**, Plot shows correspondence of PICS SNPs to indicated functional elements, compared to random SNPs from the same loci (error bars indicate standard deviation from 1,000 iterations using locus-matched control SNPs).

likely causal SNP (>75% probability) at 12% of loci linked to autoimmunity. However, most GWAS signals could not be fully resolved due to LD and thus contain several candidate causal SNPs (Fig. 1e).

To confirm the functional significance of fine-mapped SNPs, we compared PICS SNPs against a strict background of random SNPs drawn from the same loci. Candidate causal SNPs derived by PICS were strongly enriched for protein-coding (missense, nonsense, frameshift) changes, which account for 14% of the predicted causal variants compared to just 4% of the random SNPs. Modest enrichments over the locus background were also observed for synonymous substitutions (5%), 3' UTRs (3%), and splice junctions (0.2%) (Fig. 1f). Although these results support the efficacy of PICS for identifying causal variants, ~90% of GWAS hits for autoimmune diseases remain unexplained by protein-coding variants. Candidate causal SNPs and the PICS algorithm are available through an accompanying online portal (<http://www.broadinstitute.org/pubs/finemapping>).

Causal SNPs map to immune enhancers

To investigate the functions of predicted causal non-coding variants, we generated a resource of epigenomic maps for specialized immune subsets (Extended Data Fig. 6). We examined primary human CD4⁺ T-cell populations from pooled healthy donor blood, including FOXP3⁺ CD25^{hi} CD127^{lo} regulatory (T_{regs}), CD25⁺ CD45RA⁺ CD45RO⁺ naive (T_{naive}) and CD25⁺ CD45RA⁺ CD45RO⁺ memory (T_{mem}) T cells, and *ex vivo* phorbol myristate acetate (PMA)/ionomycin stimulated CD4⁺ T cells separated into IL-17-positive (CD25⁺ IL17A⁺; T_H17) and IL-17-negative (CD25⁺ IL17A⁺; T_Hstim) subsets. We also examined naive and memory CD8⁺ T cells, B cell centroblasts from paediatric tonsils (CD20⁺ CD10⁺ CXCR4⁺ CD44⁺), and peripheral blood B cells (CD20⁺) and monocytes (CD14⁺). We mapped six histone modifications by

chromatin immunoprecipitation followed by sequencing (ChIP-seq) for all ten populations, and performed RNA sequencing (RNA-seq) for each CD4⁺ T-cell population. We also incorporated data for B lymphoblastoid cells¹⁷, TH0, TH1 and TH2 stimulated T cells¹⁰, and non-immune cells from the NIH Epigenomics Project²⁵ and ENCODE²⁶, for a total of 56 cell types.

For each cell type, we computed a genome-wide map of *cis*-regulatory elements based on H3 lysine 27 acetylation (H3K27ac), a marker of active promoters and enhancers¹². We then clustered cell types based on these *cis*-regulatory element patterns (Extended Data Fig. 7). Fine distinctions could be drawn between CD4⁺ T-cell subsets based on quantitative differences in H3K27ac at thousands of putative enhancers (Fig. 2a). These cell-type-specific H3K27ac patterns correlate with the expression of proximal genes. In contrast, H3 lysine 4 mono-methylation (H3K4me1) was more uniform across subsets, consistent with its association to open or 'poised' sites shared between related cell types¹².

Mapping of autoimmune disease PICS SNPs to these regulatory annotations revealed enrichment in B-cell and T-cell enhancers (Fig. 2a). A disproportionate correspondence to enhancers activated upon T-cell stimulation prompted us to examine such elements more closely. Substantial subsets of immune-specific enhancers markedly increase their H3K27ac signals upon *ex vivo* stimulation, often in conjunction with non-coding eRNA transcription, and induction of proximal genes (Fig. 2a, b). Compared to naive T cells, enhancers in stimulated T cells are strongly enriched for consensus motifs recognized by AP-1 transcription factors, master regulators of cellular responses to stimuli. PICS SNPs are strongly enriched within stimulus-dependent enhancers ($P < 10^{-20}$ for combined PMA/ionomycin; $P < 10^{-11}$ for combined CD3/CD28), whereas enhancers preferentially marked in unstimulated T cells show no enrichment for causal variants. Candidate causal SNPs were further

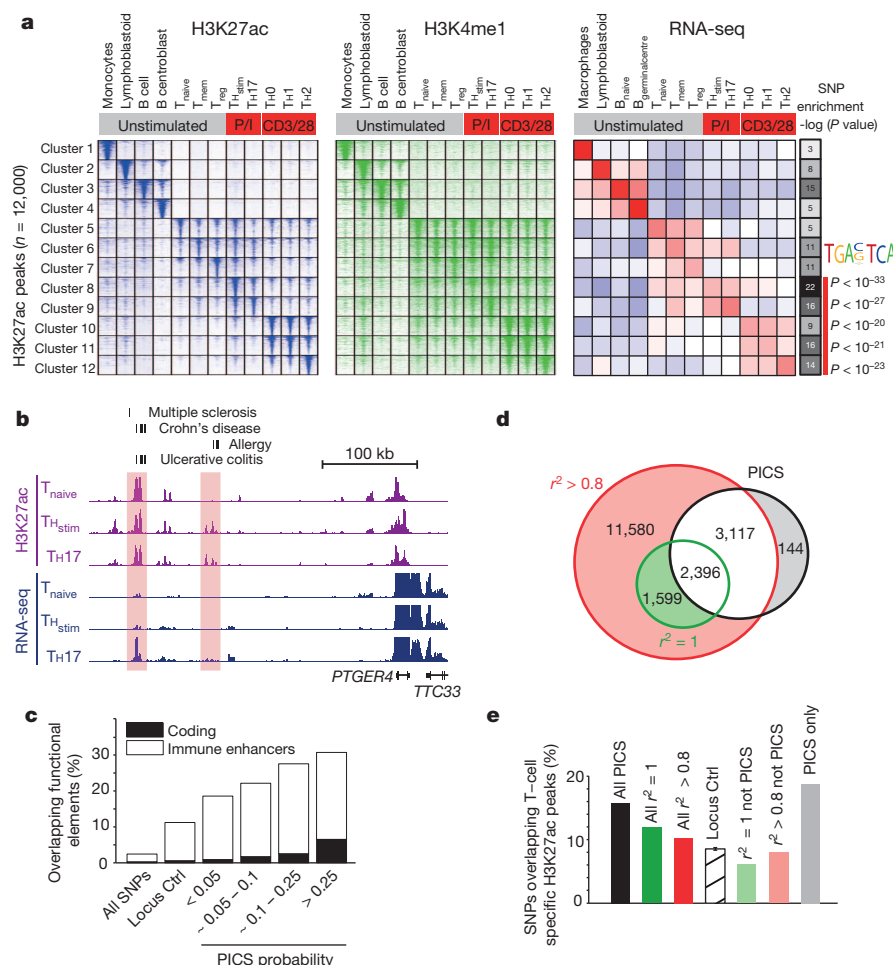


Figure 2 | Epigenetic fine mapping of enhancers. **a**, Heatmaps show H3K27ac and H3K4me1 signals for 1,000 candidate enhancers (rows) in 12 immune cell types (columns). Enhancers are clustered by the cell type-specificity of their H3K27ac signals. Adjacent heatmap shows average RNA-seq expression for the genes nearest to the enhancers in each cluster. Greyscale (right) depicts the enrichment of PICS autoimmunity SNPs in each enhancer cluster (hypergeometric P values calculated based on the number of PICS SNPs overlapping enhancers from each cluster, relative to random SNPs from the same loci). The AP-1 motif is over-represented in enhancers preferentially marked in stimulated T cells, compared to naive T cells. **b**, Candidate causal SNPs displayed along with H3K27ac and RNA-seq signals at the *PTGER4* locus. A subset of enhancers with disease variants (shaded) shows evidence of stimulus-dependent eRNA transcription. **c**, Stacked bar graph indicates percentage overlap with immune enhancers and coding sequence for PICS SNPs at different probability thresholds, compared to control SNPs drawn from the entire genome (all SNPs) or the same loci (locus Ctrl). **d**, Venn diagram compares PICS SNPs to GWAS catalogue SNPs with indicated r^2 thresholds. **e**, Bar graph indicates percentage overlap with annotated T-cell enhancers for PICS SNPs, GWAS catalogue SNPs at indicated r^2 thresholds, locus control SNPs, and three subsets of SNPs defined and shaded as in panel d.

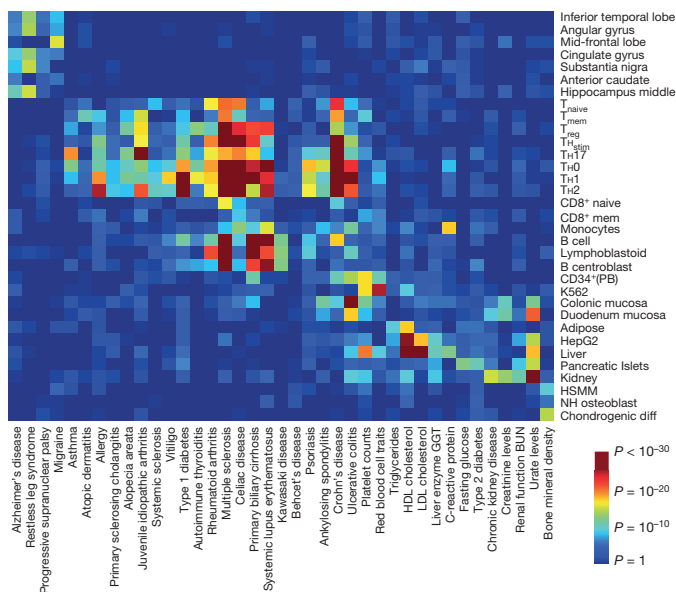


Figure 3 | Cell-type specificity of human diseases. Heatmap depicts enrichment (red = high; blue = low) of PICS SNPs for 39 diseases/traits in acetylated *cis*-regulatory elements of 33 different cell types.

enriched in T-cell enhancers that produce non-coding RNAs upon stimulation (1.6-fold; $P < 0.01$).

The association of candidate causal SNPs to immune enhancers increases with PICS probability score (Fig. 2c). We estimate that immune enhancers overall account for ~60% of candidate causal SNPs, whereas promoters account for another ~8% of these variants (Extended Data Fig. 7). When we compared these statistics against GWAS catalogue SNPs, which were the focus of prior studies linking GWAS to regulatory annotations^{10,16–19,21}, we found that the subset of associated SNPs that do not correspond to a PICS SNP fail to show any enrichment for T-cell enhancers, relative to locus controls (Fig. 2d, e). These data support the efficacy of PICS and link probable causal autoimmune disease variants to specific enhancers activated upon immune stimulation.

Cell-type signatures of complex diseases

Along with the 21 autoimmune diseases, we predicted causal SNPs for 18 other traits and diseases (Methods). Comparing SNP locations with chromatin maps for 56 cell types revealed the cell-type specificities of

cis-regulatory elements that coincide with PICS SNPs, thus predicting cell types contributing to each phenotype (Fig. 3). The patterns are more informative than the expression patterns of genes targeted by coding GWAS hits (Extended Data Fig. 8). Notable examples include SNPs associated with Alzheimer's disease and migraine, which map to enhancers and promoters active in brain tissues, and SNPs associated with fasting blood glucose, which map to elements active in pancreatic islets. Nearly all of the autoimmune diseases preferentially mapped to enhancers and promoters active in CD4⁺ T-cell subpopulations. However, a few diseases, such as systemic lupus erythematosus, Kawasaki disease, and primary biliary cirrhosis, preferentially mapped to B-cell elements. Notably, ulcerative colitis also mapped to gastrointestinal tract elements, consistent with its bowel pathology. Although the primary signature of type 1 diabetes SNPs is in T-cell enhancers, there is also enrichment in pancreatic islet enhancers ($P < 10^{-7}$). Thus, although immune cell effects may be shared among autoimmune diseases, genetic variants affecting target organs such as bowel and pancreatic islets may shape disease-specific pathology.

Discrete functional units in super-enhancers

Genomic loci that encode cellular identity genes frequently contain large regions with clustered or contiguous enhancers bound by transcriptional co-activators and marked by H3K27ac. Recent studies showed that such 'super-enhancer' regions are enriched for GWAS catalogue SNPs, including those related to autoimmunity^{18,19}. Consistently, we find that PICS SNPs are 7.5-fold enriched in CD4⁺ T-cell super-enhancers, relative to random SNPs from the genome. We therefore parsed the topography of super-enhancers in immune cells using our genetic and epigenetic data.

The *IL2RA* locus exemplifies the complex landscape of enhancer regulation. *IL2RA* encodes a receptor with key roles in T-cell stimulation and T_{reg} function¹⁵. The super-enhancer in this locus comprises a cluster of elements recognizable as distinct H3K27ac peaks (Fig. 4a). Although the region meets the super-enhancer definition in multiple CD4⁺ T-cell types¹⁸, sub-elements are preferentially acetylated in T_{reg}, TH17 and/or TH_{stim} T-cells, consistent with differential regulation. Some sub-elements appear bound by T-cell master regulators, including FOXP3 in T_{regs}, T-bet (also known as TBX21) in TH1 cells, and GATA3 in TH2 cells. A systematic analysis indicates PICS SNPs are most enriched at distinct stimulus-dependent H3K27ac peaks within super-enhancer regions (Extended Data Fig. 7).

PICS SNPs for eight autoimmune diseases map to distinct segments of the *IL2RA* super-enhancer. For example, Immunochip data identify a candidate causal SNP for multiple sclerosis that has no effect on autoimmune thyroiditis disease risk. Conversely, a candidate causal SNP for

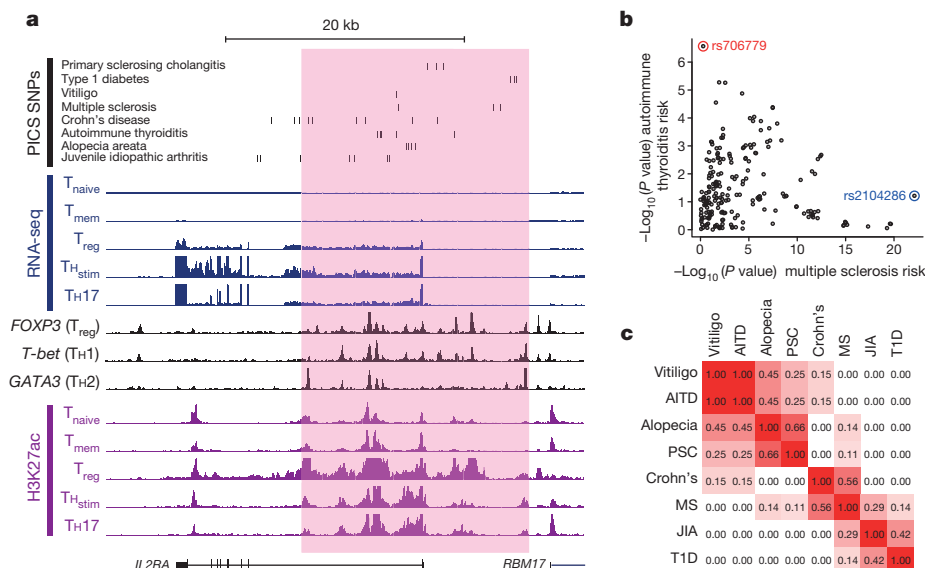


Figure 4 | Disease variants map to discrete elements in super-enhancers. **a**, Candidate causal SNPs for autoimmune diseases are displayed along with H3K27ac, RNA-seq and transcription factor binding profiles for the *IL2RA* locus, which contains a super-enhancer (pink shade). **b**, For all SNPs in the *IL2RA* locus, scatter plot compares strength of association with multiple sclerosis versus autoimmune thyroiditis. Immunochip data resolve rs706779 (red) as the lead SNP for autoimmune thyroiditis and rs2104286 (blue) as the lead SNP for multiple sclerosis. **c**, LD matrix displaying r^2 between lead SNPs for different diseases at the *IL2RA* locus confirms distinct and independent genetic associations within the super-enhancer. AITD, autoimmune thyroiditis; JIA, juvenile idiopathic arthritis; MS, multiple sclerosis; PSC, primary sclerosing cholangitis; T1D, type 1 diabetes.

autoimmune thyroiditis has no effect on multiple sclerosis risk, despite the proximity of the two SNPs within the super-enhancer (Fig. 4b). Furthermore, index SNPs for multiple other diseases are not in LD, suggesting that multiple sites of nucleotide variation in the locus have separable disease associations (Fig. 4c). The distribution of PICS SNPs and the partially discordant regulation of sub-regions suggest that super-enhancers may comprise multiple discrete units with distinct regulatory signals, functions, and phenotypic associations.

Disease SNPs fall near consensus motifs

The enrichment of candidate causal variants within enhancers suggests that they affect disease risk by altering gene regulation, but does not distinguish the underlying mechanisms. Enhancer activity is dependent on complex interplay between transcription factors, chromatin, non-coding RNAs and tertiary interactions of DNA loci²⁷. A straightforward hypothesis is that disease SNPs alter transcription factor binding. Indeed, PICS SNPs tend to coincide with nucleosome-depleted regions, characterized by DNase hypersensitivity and localized (~150 bp) dips in H3K27ac signal²⁶, which are indicative of transcription factor occupancy (Fig. 5a).

We therefore overlapped PICS SNPs with 31 transcription factor binding maps generated by ENCODE²⁶ (Fig. 5b). Candidate causal SNPs are strongly enriched within binding sites for immune-related transcription factors, including NF- κ B, PU1 (also known as SPI1), IRF4, and BATF. Variants associated with different diseases correlate to different combinations of transcription factors that control immune cell identity and response to stimulation. For example, multiple sclerosis SNPs preferentially coincide with NF- κ B, EBF1 and MEF2A-bound regions, whereas rheumatoid arthritis and coeliac disease SNPs preferentially coincide with IRF4 regions.

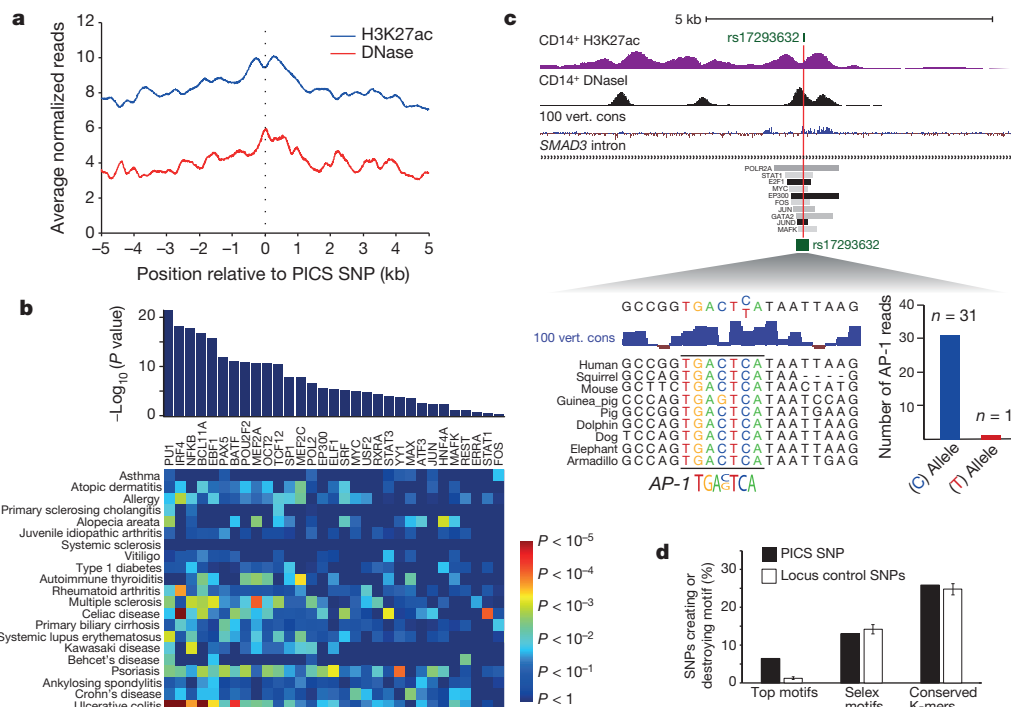


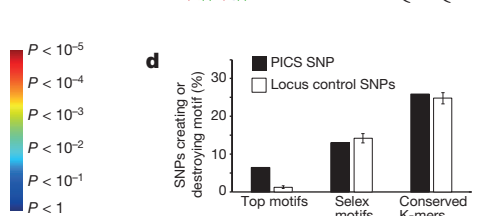
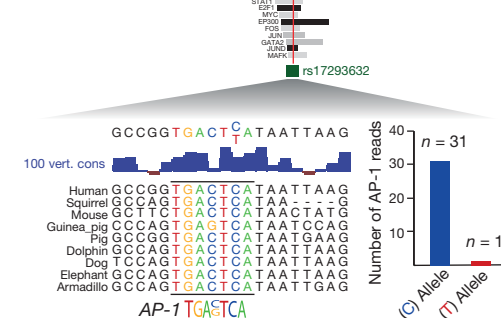
Figure 5 | Causal variants map to regions of transcription factor binding. **a**, Plot depicts composite H3K27ac and DNase signals²⁶ in immune cells over PICS autoimmunity SNPs. Overall PICS SNPs coincide with nucleosome-depleted, hypersensitive sites, indicative of transcription factor binding. **b**, Bar plot indicates transcription factors whose binding is enriched near PICS SNPs for all 21 autoimmune diseases²⁶. Heatmap depicts enrichment of these transcription factors near variants associated with specific diseases (red = high; blue = low). **c**, H3K27ac, DNase²⁶ and conservation signals, and selected transcription factor binding intervals are shown in a *SMAD3* intronic locus. rs17293632, a non-coding candidate causal SNP for Crohn's disease, disrupts a

conserved AP-1 binding motif in an enhancer marked by H3K27ac in CD14⁺ monocytes. Summing of ChIP-seq reads overlapping the SNP in the heterozygous HeLa cell line shows that only the intact motif binds AP-1 transcription factors, Jun and Fos. **d**, Bar graph shows the fraction of PICS SNPs (black) versus random SNPs from the same locus (white) that create or disrupt one of the significantly enriched motifs, any SELEX (systematic evolution of ligands by exponential enrichment) motif, or any conserved K-mer. Error bars indicate standard deviation from 1,000 iterations using locus-matched control SNPs.

Next, we examined whether causal variants disrupt or create cognate sequence motifs recognized by these transcription factors. We focused on 823 of the highest-likelihood non-coding PICS SNPs, an estimated 30% of which represent true causal variants. We identified PICS SNPs that alter motifs for NF- κ B ($n = 2$), AP-1 ($n = 8$), or ETS/ELF1 ($n = 5$). Overall, we identified 7 known transcription factor motifs and 6 conserved sequence motifs^{28,29} with a significant tendency to overlap causal variants likely to alter binding affinity. Of the highest-likelihood SNPs, 7% affected one of these over-represented motifs, with a roughly equal distribution between motif creation and disruption (Extended Data Fig. 9).

A notable motif-disrupting PICS SNP is the Crohn's disease-associated variant rs17293632 (C > T, minor allele increases disease risk; PICS probability ~54%), which resides in an intron of *SMAD3* (Fig. 5c). *SMAD3* encodes a transcription factor downstream of transforming growth factor β (TGF- β) with pleiotropic roles in immune homeostasis³⁰. The SNP disrupts a conserved AP-1 consensus site. ChIP-seq data for AP-1 transcription factors (Jun, Fos) in a heterozygous cell line reveal robust binding to the reference sequence, but not to the variant sequence created by the SNP. As described above, a prominent AP-1 signature is associated with enhancers activated upon immune stimulation (Fig. 2a). This suggests that rs17293632 may increase Crohn's disease risk by directly disrupting AP-1 regulation of the TGF- β -*SMAD3* pathway.

Despite this and other compelling examples, only ~7% of the highest-likelihood non-coding PICS SNPs alter an over-represented transcription factor motif. Scanning a large database of transcription factor motifs, we found that ~13% of high-likelihood causal SNPs create or disrupt some known consensus sequence derived by *in vitro* selection²⁸, whereas ~27% create or disrupt a putative consensus sequence derived from phylogenetic analysis²⁹. However, these proportions are similar to the rate for background SNPs (Fig. 5d). Even extrapolating for uncertainty



conserved AP-1 binding motif in an enhancer marked by H3K27ac in CD14⁺ monocytes. Summing of ChIP-seq reads overlapping the SNP in the heterozygous HeLa cell line shows that only the intact motif binds AP-1 transcription factors, Jun and Fos. **d**, Bar graph shows the fraction of PICS SNPs (black) versus random SNPs from the same locus (white) that create or disrupt one of the significantly enriched motifs, any SELEX (systematic evolution of ligands by exponential enrichment) motif, or any conserved K-mer. Error bars indicate standard deviation from 1,000 iterations using locus-matched control SNPs.

in causal SNP assignments, our data suggest that at most 10–20% of non-coding GWAS hits act by altering a recognizable transcription factor motif.

Notwithstanding their infrequent coincidence to the precise transcription factor motifs, non-coding PICS SNPs have a strong tendency to reside in close proximity to such sequences. Candidate causal variants are most significantly enriched in the vicinity of NF- κ B, RUNX1, AP-1, ELF1, and PU1 motifs (Extended Data Fig. 9), with 26% residing within 100 bp of such a motif. These findings parallel recent studies of genetic variation in mice, where DNA variants affecting NF- κ B binding are dispersed in the vicinity of the actual binding sites³¹. Our results suggest that many causal non-coding SNPs modulate transcription factor dependent enhancer activity (and confer disease risk) by altering adjacent DNA bases whose mechanistic roles are not readily explained by existing gene regulatory models.

Gene regulatory effects of disease SNPs

To assess the effects of autoimmunity-associated genetic variation on gene regulation, we incorporated a recent study that mapped variants associated with heritable differences in peripheral blood gene expression³². We used PICS to predict causal expression quantitative locus (eQTL) SNPs, which we compared against random SNPs from the same loci. These eQTL SNPs are strongly enriched in promoters (9%) and 3' UTRs (25%), but show relatively modest preference for immune enhancers (14%), compared to GWAS SNPs (Fig. 6a). Overall, ~12% of causal non-coding autoimmune disease variants also score as eQTL SNPs (Extended Data Fig. 10). Disease SNPs that did not score as eQTLs in peripheral blood may score in more precise immune subsets in relevant regulatory contexts. Nonetheless, their modest overlap with eQTLs and their striking correspondence to enhancers suggest that most disease variants exert subtle and highly context-specific effects on gene regulation.

Incorporation of eQTL SNPs allowed us to link causal non-coding disease variants to specific genes. For example, PICS fine mapping identified two SNPs in the *IKZF3* locus with independent effects on *IKZF3* expression, rs12946510 and rs907091. *IKZF3* encodes an IKAROS family transcription factor with key roles in lymphocyte differentiation and function³³. Interestingly, the minor allele of rs12946510 is associated with decreased *IKZF3* expression and increased multiple sclerosis risk (Fig. 6b, c), whereas the minor allele of rs907091 is associated with increased *IKZF3* expression, but does not affect disease risk. This suggests that disease risk is dependent on the specific mode and context in which a variant influences gene expression.

Despite strong evidence from fine mapping that rs12946510 is the causal SNP affecting multiple sclerosis risk and *IKZF3* expression, the underlying sequence does not reveal a clear mechanism of action. The disease SNP resides within a conserved element with enhancer-like chromatin in immune cells. It coincides with a nucleosome-depleted, DNase hypersensitive site bound by multiple transcription factors, including immune-related factors RUNX3, RELA (NF- κ B family member), EBF1, POU2F2 and MEF2 (Fig. 6d). The C/T variation at this site does not create or disrupt a readily recognizable consensus DNA motif, but overlaps a highly degenerate MEF2 motif and might thus modulate transcription factor binding despite incomplete sequence specificity. This example illustrates the value of integrative functional genomic analysis for investigating the complex mechanisms by which non-coding variants modulate gene expression and disease risk.

Discussion

Interpretation of non-coding disease variants, which comprise the vast majority of GWAS hits, remains a momentous challenge due to haplotype structure and our limited understanding of the mechanisms and physiological contexts of non-coding elements. Here we addressed these issues through combination of high-density genotyping and epigenomic data. Focusing on autoimmune diseases, we triaged causal variants based solely on genetic evidence and integrated chromatin and transcription factor binding maps to distinguish their probable functions and physiological contexts. We found that most causal variants map to enhancers and

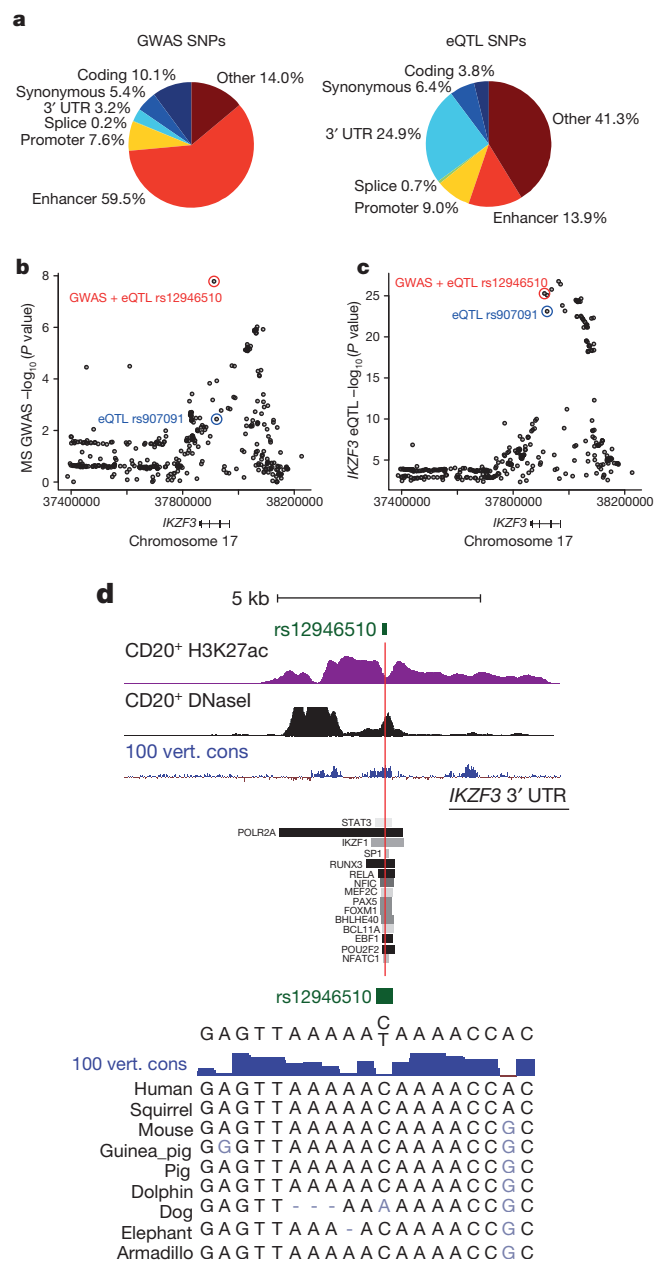


Figure 6 | Functional effects of disease variants on gene expression.

a, Pie charts show the fraction of PICS autoimmunity SNPs (left) or peripheral blood eQTLs (right) explained by the indicated genomic features. **b**, GWAS signal for multiple sclerosis risk at the *IKZF3* locus. The minor allele of rs12946510 (red) is associated with both disease risk and eQTL effect (decreased *IKZF3* expression), while the minor allele of rs907091 (blue) scored as an eQTL only (increased *IKZF3* expression). **c**, eQTL association signal for *IKZF3* shown for the same regions as in **b**. **d**, H3K27ac, DNaseI and conservation signals, and selected transcription factor binding intervals are shown in the vicinity of rs12946510, which occurs in a conserved site marked by H3K27ac in multiple cell types, including CD20⁺ B cells, and bound by multiple transcription factors. The C/T variation at this SNP does not disrupt any clearly defined DNA motif, but coincides with a degenerate MEF2 motif.

frequently coincide with nucleosome-depleted sites bound by immune-related transcription factors. The resulting resource highlights specific transcription factors, target loci and pathways with disease-specific or general roles in autoimmunity.

Yet despite their close proximity to immune transcription factor binding sites, only a fraction of causal non-coding variants alter recognizable transcription factor sequence motifs. Moreover, disease variants have a distinct functional distribution and infrequently overlap peripheral

blood eQTLs, which suggests that they exert highly contextual regulatory effects. Although these features of non-coding disease variants further challenge GWAS interpretation, they might not be unexpected. Biochemical and genetic manipulations have established the potential of motif-adjacent sequences to influence transcription factor activity³⁴. Roles for such non-canonical sequences are also supported by the extended nucleotide conservation at many enhancers, most of which lies outside of known motifs, and the complex structural interactions and looping events that underlie gene regulation²⁷. Furthermore, common variants contributing to polygenic autoimmunity are expected to have modest, context-restricted effects, given that strongly deleterious mutations would be eliminated from the population¹. Compared to mutations that disrupt transcription factor motifs, alterations to non-canonical determinants may produce subtle but pivotal alterations to the immune response, without reaching a level of disruption that would result in strong negative selection.

Systematic integration of fine-mapped genetic and epigenetic data implies a nuanced complexity to disease variant function that will continue to push the limits of experimental and computational approaches. Much work remains to be done to characterize SNPs whose causality can be firmly established through genotyping and to facilitate efforts to resolve GWAS signals that remain refractory to fine mapping due to haplotype structure. Understanding their regulatory mechanisms could have broad implications for autoimmune disease biology and treatment, given genetic links to immune regulators, such as NF- κ B, IL2RA and IKZF3 (also known as AIOLOS), and implied transcriptional and epigenetic aberrations, all of which are candidates for therapeutic intervention.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 11 February; accepted 4 September 2014.

Published online 29 October 2014.

- Altshuler, D., Daly, M. J. & Lander, E. S. Genetic mapping in human disease. *Science* **322**, 881–888 (2008).
- Hindorf, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA* **106**, 9362–9367 (2009).
- Vyse, T. J. & Todd, J. A. Genetic analysis of autoimmune disease. *Cell* **85**, 311–318 (1996).
- Buckner, J. H. Mechanisms of impaired regulation by CD4⁺CD25⁺FOXP3⁺ regulatory T cells in human autoimmune diseases. *Nature Rev. Immunol.* **10**, 849–859 (2010).
- Browning, J. L. B cells move to centre stage: novel opportunities for autoimmune disease treatment. *Nature Rev. Drug Discov.* **5**, 564–576 (2006).
- Zhou, L., Chong, M. M. & Littman, D. R. Plasticity of CD4⁺ T cell lineage differentiation. *Immunity* **30**, 646–655 (2009).
- Ciofani, M. *et al.* A validated regulatory network for Th17 cell specification. *Cell* **151**, 289–303 (2012).
- Marson, A. *et al.* Foxp3 occupancy and regulation of key target genes during T-cell stimulation. *Nature* **445**, 931–935 (2007).
- Samstein, R. M. *et al.* Foxp3 exploits a pre-existent enhancer landscape for regulatory T cell lineage specification. *Cell* **151**, 153–166 (2012).
- Hawkins, R. D. *et al.* Global chromatin state analysis reveals lineage-specific enhancers during the initiation of human T helper 1 and T helper 2 cell polarization. *Immunity* **38**, 1271–1284 (2013).
- Vahedi, G. *et al.* STATs shape the active enhancer landscape of T cell populations. *Cell* **151**, 981–993 (2012).
- Rivera, C. M. & Ren, B. Mapping human epigenomes. *Cell* **155**, 39–55 (2013).
- Ostuni, R. *et al.* Latent enhancers activated by stimulation in differentiated cells. *Cell* **152**, 157–171 (2013).
- Lam, M. T. *et al.* Rev-Erbs repress macrophage gene expression by inhibiting enhancer-directed transcription. *Nature* **498**, 511–515 (2013).
- Parkes, M., Cortes, A., van Heel, D. A. & Brown, M. A. Genetic insights into common pathways and complex relationships among immune-mediated diseases. *Nature Rev. Genet.* **14**, 661–673 (2013).
- Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
- Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).
- Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934–947 (2013).
- Parker, S. C. *et al.* Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc. Natl Acad. Sci. USA* **110**, 17921–17926 (2013).
- International Multiple Sclerosis Genetics Consortium *et al.* Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nature Genet.* **45**, 1353–1360 (2013).
- Trynka, G. *et al.* Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nature Genet.* **45**, 124–130 (2013).
- 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
- West, L. C. & Cresswell, P. Expanding roles for GILT in immunity. *Curr. Opin. Immunol.* **25**, 103–108 (2013).
- International Multiple Sclerosis Genetics Consortium & The Wellcome Trust Case Consortium 2. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature* **476**, 214–219 (2011).
- Bernstein, B. E. *et al.* The NIH roadmap epigenomics mapping consortium. *Nature Biotechnol.* **28**, 1045–1048 (2010).
- The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
- Bulger, M. & Groudine, M. Functional and mechanistic diversity of distal transcription enhancers. *Cell* **144**, 327–339 (2011).
- Jolma, A. *et al.* DNA-binding specificities of human transcription factors. *Cell* **152**, 327–339 (2013).
- Xie, X. *et al.* Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* **434**, 338–345 (2005).
- Li, M. O. & Flavell, R. A. TGF- β : a master of all T cell trades. *Cell* **134**, 392–404 (2008).
- Heinz, S. *et al.* Effect of natural genetic variation on enhancer selection and function. *Nature* **503**, 487–492 (2013).
- Wright, F. A. *et al.* Heritability and genomics of gene expression in peripheral blood. *Nature Genet.* **46**, 430–437 (2014).
- Quintana, F. J. *et al.* Aiolos promotes Th17 differentiation by directly silencing *Il2* expression. *Nature Immunol.* **13**, 770–777 (2012).
- Gordan, R. *et al.* Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. *Cell Reports* **3**, 1093–1104 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank members of the NIH Epigenomics Consortium, M. Greenberg, H. Chang and G. Haliburton for constructive comments. We also thank IIBDGC and P. Sullivan for sharing data pre-publication, and G. Cvetanovich, S. Bhela, C. Hartnick, F. Pfeffer, D. Dombkowski and the Brigham and Women's Hospital PhenoGenetic Project for assistance with data collection. This research was supported by the NIH Common Fund (ES017155), the National Human Genome Research Institute (HG004570), the National Institute of Allergy and Infectious Disease (AI045757, AI046130, AI070352, AI039671), the National Institute of Neurological Disorders and Stroke (NS24247, NS067305), the National Institute of General Medical Sciences (GM093080), the National Multiple Sclerosis Society (CA1061-A-18), the UCSF Sandler Fellowship, a gift from Jake Aronov, the Penates Foundation, the Nancy Taylor Foundation, and the Howard Hughes Medical Institute.

Author Contributions A.M., D.A.H. and B.E.B. designed the study. K.K.F. performed genetic analysis, PICS development and integration. M.J.D. supervised genetic analysis. J.Z., M.K., W.J.H., S.B., N.S., H.W., R.J.H.R., A.A.S., M.H., M.J.C.-A., D.M., C.J.L., V.K.K. and C.B.E. contributed to data collection and analysis. N.A.P. and P.L.D.J. contributed multiple sclerosis genotyping data. K.K.F., A.M., D.A.H. and B.E.B. wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.M. (alexander.marson@ucsf.edu).

METHODS

Cell isolation and culture

Purification and culture of human CD4⁺ T-cell subsets. Cells were obtained from the peripheral blood of pooled healthy subjects in compliance with Institutional Review Board (Yale University and Partners Human Research Committee) protocols. Untouched CD4⁺ T cells were isolated by gradient centrifugation (Ficoll-Hypaque; GE Healthcare) using the RosetteSep Human CD4⁺ T-cell Enrichment kit (StemCell Technologies). CD4⁺ T cells were next subjected to anti-CD25 magnetic bead labelling (Miltenyi Biotec), to allow magnetic cell separation (MACS) of CD25⁺ and CD25⁻ cells. Subsequently CD25⁺ cells were stained with fluorescence-labelled monoclonal antibodies to CD4, CD25 and CD127 (BD Pharmingen), and sorted using a FACS ARIA (BD Biosciences) for CD25^{hi}CD127^{lo/-} T_{reg} cells, which express FOXP3 (Biolegend) as confirmed by intracellular post-sort analysis by FACS (Extended Data Fig. 6). Dead cells were excluded by propidium iodide (BD). An aliquot of CD25⁻ cells was labelled with fluorescence-labelled monoclonal antibodies to CD4, CD45RA and CD45RO (BD Pharmingen), and sorted on a FACS ARIA to isolate CD45RO⁺CD45RA⁻ memory (T_{mem}) and CD45RO⁻CD45RA⁺ naive (T_{naive}) CD4⁺ T-cell populations. Dead cells were excluded by propidium iodide. Highly pure human Th17 cells were isolated with modifications as previously described³⁵. In brief, CD25⁻ cells were stimulated in serum-free X-VIVO15 medium (BioWhittaker) with PMA (50 ng ml⁻¹) and ionomycin (250 ng ml⁻¹; both from Sigma-Aldrich) for 8 h and sorted by a combined MACS and FACS cell sorting strategy based on surface expression of IL-17A. Stimulated cells were stained with anti-IL-17A-PE (Miltenyi) and labelled with anti-PE microbeads (Miltenyi) and subsequently pre-enriched over an LS column (Miltenyi). The IL-17A negative fraction was used as control population (Th_{stim}). MACS-enriched Th17 cells were further sorted on a FACS ARIA (BD) for highly pure IL-17A⁺ cells (Th17).

Purification of human naive and memory CD8⁺ T cells. Leukocyte-enriched fractions of peripheral blood (byproduct of Trima platelet collection) from anonymous healthy donors were obtained from the Kraft Family Blood Donor Center (DFCI, Boston, MA) in compliance with the institutional Investigational Review Board (Partners Human Research Committee) protocol. For two independent purifications of each cell subset, blood fractions from 7 and 8 donors were pooled. Total T cells were isolated by immunodensity negative selection using the RosetteSep Human T-cell Enrichment Cocktail (STEMCELL Technologies, Vancouver, Canada) and gradient centrifugation on Ficoll-Paque PLUS (GE Healthcare, Pittsburgh, PA), according to the manufacturer's instructions. Subsequently, T cells were stained at 4 °C for 30 min using fluorescently labelled monoclonal anti-human CD8 (FITC, 2.5 µg ml⁻¹, clone RPA-T8, Biolegend, San Diego, CA), CD4 (PE, 1.25 µg ml⁻¹, clone RPA-T4, Biolegend), CD45RA (PerCP-Cy5.5, 2.4 µg ml⁻¹, clone HI100, eBioscience, San Diego, CA) and CD45RO (APC, 0.6 µg ml⁻¹, clone UCHL1, eBioscience) antibodies diluted in staining buffer (PBS supplemented with 2% fetal bovine serum, FBS), 4',6-diamidino-2-phenylindole (DAPI, 2.5 µg ml⁻¹, Life Technologies, Grand Island, NY) was also included to stain for dead cells. After washing with staining buffer, naive (CD45RA⁺CD45RO⁻) and memory (CD45RA⁻CD45RO⁺) CD8⁺ or CD4⁺ were isolated using a BD FACSAria 4-way cell sorter (BD Biosciences, San Jose, CA). Cell subsets were identified using a BD FACSDiva Software (BD Biosciences) after gating on lymphocytes (by plotting forward versus side scatters) and excluding aggregated (by plotting forward scatter pulse height versus pulse area), dead (DAPI⁺), and CD8/CD4 double positive cells (Extended Data Fig. 6). Cell purity was 90–94% CD8⁺ or 97–99% CD4⁺, and > 99% naive or memory.

Purification of human B centroblasts. Cells were obtained in compliance with Institutional Review Board (Partners Human Research Committee) protocols. For purification of human centroblasts, bulk mononuclear cells were isolated from fresh paediatric tonsillectomy specimens by mechanical disaggregation and Ficoll-Paque centrifugation³⁶. MACS enrichment of germinal centre cells was performed using anti-CD10-PE-Cy7 (BD Biosciences), and anti-PE microbeads (Miltenyi Biotec). Centroblasts³⁷ (CD19⁺CD10⁺CXCR4⁺CD44⁺CD3⁻) were purified from the enriched germinal centre cells by FACS antibodies for CD19 (APC, clone SJ25C1, BD), CD3 (BV606, clone OKT3, Biolegend), CD10 (PE-Cy7, clone HI10A, BD), CD44 (FITC, clone LI78, BD) and CXCR4 (PE, clone 12G5, eBioscience) (Extended Data Fig. 6).

Purification of adult human peripheral B cells and monocytes. Human peripheral B cells and monocytes were provided by the S. Heimfeld laboratory at the Fred Hutchinson Cancer Research Center. The cells were obtained from human leukapheresis product using standard procedures. Briefly, peripheral B cells (CD20⁺CD19⁺) and monocytes (CD14⁺) were isolated by immunomagnetic separation using the CliniMACS affinity-based technology (Miltenyi Biotec GmbH, Bergisch Gladbach, Germany) according to the manufacturer's recommendation. Reagents, tubing sets, and buffers were purchased from Miltenyi Biotec.

ChIP-seq. Following isolation (\pm *ex vivo* stimulation), cells were crosslinked in 1% formaldehyde at room temperature or 37 °C for 10 min in preparation for ChIP. Chromatin immunoprecipitation and sequencing were performed as previously described³⁸. Data sets were publicly released upon verification at (<http://epigenomeatlas.org>).

RNA-seq. RNA was extracted from CD4⁺ T-cell subsets with TRIzol. Briefly, polyadenylated RNA was isolated using oligo dT beads (Invitrogen) and fragmented to 200–600 base pairs and then ligated to RNA adaptors using T4 RNA ligase (NEB), preserving strand of origin information as previously described^{39,40}.

Enhancer annotation and clustering. ChIP-seq data were processed as previously described³⁸. Briefly, ChIP-seq reads of 36 bp were aligned to the reference genome (hg19) using the Burroughs–Wheeler Alignment tool (BWA)⁴¹. Reads aligned to the same position and strand were only counted once. Aligned reads were extended by 250 bp to approximate fragment sizes and then a 25-bp resolution chromatin map was derived by counting the number of fragments overlapping each position. H3K27ac and H3K4me1 peaks were identified by scanning the genome for enriched 1 kb windows and then merging all enriched windows within 1 kb, using as a threshold 4 genome-normalized reads per base pair³⁸. Adjacent windows separated by gaps less than 500 bp in size were joined. H3K27ac peaks that do not overlap a \pm 2.5 kb region of an annotated transcriptional start site (TSS) were defined as candidate distal regulatory elements. In order to define the cell-specific H3K27ac peaks, we calculated the mean signal in 5 kb regions centred at distal H3K27ac peaks and sorted the peaks by the ratio of signal in one cell type to all remaining cell types. For each immune cell type, the top 1,000 distal H3K27ac peaks with highest ratio were catalogued as the cell-specific distal H3K27ac peaks (Fig. 2). The heatmaps for H3K27ac and H3K4me1 signal were plotted over 10 kb regions surrounding all distal cell-specific H3K27ac peaks.

The distal H3K27ac peaks were assigned to their potential target genes if they locate in the gene body or within 100 kb regions upstream the TSS. Expression levels of the target genes were derived from RNA-seq data. Paired-end RNA-seq reads were aligned to RefSeq transcripts using Bowtie2 (ref. 42). RNA-seq data for B cells, B centroblast, macrophages, Th1, Th2 and Th0 were retrieved from NCBI GEO and SRA database (B_{naive}: GSE45982; B_{germinalcenter}: GSE45982 (ref. 43); Macrophages: GSE36952 (ref. 44); Th0, Th1 and Th2: SRA082670 (ref. 10)). RNA-seq data for lymphoblastoid (GM12878) was retrieved from ENCODE project²⁶. The number of reads per kilobase per million reads (RPKM) was calculated for each gene locus. Heatmap of RNA-seq data shows the average relative expression of all potential target genes for each cluster of cell type-specific regulatory elements.

Shared genetic loci for common human diseases. Publicly available GWAS catalogue data were obtained from the NHGRI website, (<http://www.genome.gov/gwastudies/>), current as of July 2013 (refs 45, 46). Studies were included based on the criteria that they had at least 6 hits at the genome-wide significant level of $P \leq 5 \times 10^{-8}$. From a set of 21 autoimmune diseases and 18 representative non-autoimmune diseases/traits, we included index SNPs with significance $P \leq 10^{-6}$ for downstream analysis.

In some cases, the same disease had multiple index SNPs mapping to the same locus (defined as within 500 kb of each other), due to independently conducted GWAS studies identifying different lead SNPs within the same region. For these loci, only the most significant GWAS index SNP was kept for downstream analysis, resulting in 1,170 GWAS index SNPs for 39 diseases/traits. For each pair of diseases/traits, we compared their respective lists of index SNPs to find instances of common genetic loci (defined as the two diseases sharing index SNPs within 500 kb of each other). The number of overlapping loci was calculated for each disease pair. To measure the genetic similarity between two diseases/traits, a disease-by-disease correlation matrix was calculated based on the number of overlapping loci for each disease/trait with each of the other diseases, and the results are shown in Fig. 1a.

Sources of Immunochip and Non-Immunochip GWAS data. Summary statistics for published Immunochip studies of coeliac disease⁴⁷, autoimmune thyroiditis⁴⁸, primary biliary cirrhosis⁴⁹, and rheumatoid arthritis⁵⁰ were downloaded from the Immunobase website, (<http://www.immunobase.org/>). Full genotype data and PCA analysis for the multiple sclerosis Immunochip GWAS study²⁰ was provided by the International Multiple Sclerosis Genetics Consortium. For ankylosing spondylitis⁵¹, atopic dermatitis⁵², primary sclerosing cholangitis⁵³, juvenile idiopathic arthritis⁵⁴, and psoriasis⁵⁵, Immunochip studies had been previously been published, but only the lead SNPs from associated Immunochip regions were available. We also included GWAS of autoimmune diseases that had not been studied using Immunochip, including asthma, allergy, Kawasaki disease, Behcet's disease, vitiligo, alopecia areata, systemic lupus erythematosus, systemic sclerosis, type 1 diabetes, Crohn's disease, and ulcerative colitis. For these diseases and the 18 representative non-immune diseases, index SNPs from the GWAS catalogue were used⁴⁶. In addition, full genotype data and PCA analysis for the inflammatory bowel disease Immunochip GWAS study were provided by the International Inflammatory Bowel Diseases Genetics Consortium for purposes of calculating the statistical models used in PICS. Because the results for the IBD Immunochip analysis are unpublished, we used the previously published index SNP results for inflammatory bowel disease from the GWAS catalogue.

Probabilistic identification of causal SNPs (PICS). We developed a fine-mapping algorithm, which we call probabilistic identification of causal SNPs (PICS), that makes use of densely-mapped genotyping data to estimate each SNP's probability of being a causal variant, given the observed pattern of association at the locus. We developed PICS on large multiple sclerosis (MS) (14,277 cases, 23,605 controls²⁰) and inflammatory bowel disease (IBD) cohorts (34,594 cases, 28,999 controls; unpublished data) that were genotyped using the Immunochip, a targeted ultra-dense genotyping array with comprehensive coverage of 1000 Genomes Project SNPs²² within 186 autoimmune disease-associated loci.

Analysis of IBD risk associated with SNPs at the *IL23R* locus presents an illustrative example of the LD problem and the potential for PICS to overcome this challenge (Extended Data Fig. 1). The most strongly associated SNP is rs11209026, a loss of function missense variant that changes a conserved arginine to glutamine at amino acid position 381 (R381Q) and decreases downstream signalling through the STAT3 pathway^{56,57}. Association with IBD decreases with physical distance along the chromosome, due to rare recombination events that break up the haplotype and distinguish the causal missense mutation from other tightly linked neutral variants. These rare informative recombination events would be missed by standard genotyping arrays with probes spread thinly across the entire genome.

For neutral SNPs whose association signal is only due to being in LD with a causal SNP, the strength of association, as measured by chi-square (or log *P* value, since chi-square and log *P* value are asymptotically linear) scales linearly with their r^2 to the causal SNP. This is because strength of association is linear with r^2 by the formula for the Armitage trend test⁵⁸:

$$\chi^2 = (n-1)r^2$$

where χ^2 is the chi-square association test statistic, n is the sample size, and r^2 is the square of the correlation coefficient.

This linear trend is observed at the *IL23R* locus, consistent with a model where R381Q is the causal variant, and neutral SNPs demonstrate association signal in proportion to their LD to the causal variant (Extended Data Fig. 1). SNPs in linkage to R381Q do not perfectly fall on the expected line, due to statistical fluctuations. Independent association studies for the same disease tend to nominate different SNPs within a given locus as their best association, due to statistical fluctuation pushing a different SNP to the forefront in each subsequent study^{59–62}. Note that a group of SNPs that are strongly associated to disease but are not in linkage with rs11209026 (R381Q) represent independent association signals at the locus.

Although we know from functional studies that R381Q is the likely causal variant, we sought additional statistical evidence to support R381Q as the causal variant, and to refute the null hypothesis that the prominent association of R381Q (compared to other SNPs in the haplotype) is due to chance. We simulated 1,000 permutations by fixing the association signal at R381Q, but with all other SNPs being neutral, while preserving the LD relationships between SNPs in the locus. An odds ratio of 1.2 was used rather than the approximately twofold odds ratio naturally observed at R381Q, because this was more representative of the modest association signal strengths observed at other GWAS loci. For each round of permutation, we obtained the association signal at all SNPs in the locus. Because only the association signal at R381Q is fixed, the signal at the remaining neutral SNPs in the locus are free to vary due to statistical fluctuations; four typical examples of simulated association results at the R381Q locus are shown (Extended Data Fig. 1), including two examples where the causal variant is not the most strongly associated SNP in the locus. From these 1,000 iterations, we calculated the standard deviation in the association signal for each of the SNPs in the *IL23R* locus (Extended Data Fig. 2). We show that the distribution of association signals for each SNP approximates a normal distribution, centred at the expected value based on that SNP's r^2 to the causal variant (Extended Data Fig. 2).

These permutations demonstrate that the causal variant need not be the most strongly associated SNP within the locus, due to statistical fluctuations. Rather, given the observed pattern of association at a locus, we are interested in knowing the probability of each SNP within the locus to be the causal variant. We can use Bayes' theorem to infer the probability of each SNP being the causal variant, by using information derived from the permutations. As the prior probability of each SNP to be the causal variant is equal, the SNP most likely to be the causal variant is therefore the SNP whose simulated signal most closely approximates the observed association at the locus. By performing permutations of a simulated association signal at each SNP within the locus, we can estimate the probability that the SNP could lead to the observed association at the locus.

For example, consider a two SNP example where SNP A and SNP B are in LD, and SNP A is the lead SNP in the locus (Extended Data Fig. 2). If we are interested in knowing $P(B^{\text{causal}}|A^{\text{lead}})$, that is, the probability that SNP B is the causal variant given that SNP A is the top signal in the locus, then by Bayes' theorem:

$$P(B^{\text{causal}}|A^{\text{lead}}) = P(A^{\text{lead}}|B^{\text{causal}}) \times P(A^{\text{lead}}) / P(B^{\text{causal}})$$

Where $P(A^{\text{lead}}|B^{\text{causal}})$ is the probability of SNP A being the top signal in the locus, given that SNP B is the causal variant. $P(A^{\text{lead}}|B^{\text{causal}})$ is straightforward to calculate by performing permutations with a simulated signal at SNP B, and measuring the number of permutations where SNP A emerges as the top signal in the locus despite SNP B being the actual causal variant. We have assumed that the prior probability of each SNP to be the causal variant or the lead SNP is equal, although this could be adjusted based on external information, such as functional annotation of the SNP to be a coding variant.

Using the formula above, we calculate both $P(B^{\text{causal}}|A^{\text{lead}})$ and $P(A^{\text{causal}}|A^{\text{lead}})$, and then normalize both of these probabilities so that $P(B^{\text{causal}}|A^{\text{lead}}) + P(A^{\text{causal}}|A^{\text{lead}}) = 1$. In cases where there are more than two SNPs to consider, we similarly normalize the probabilities so that they sum to 1. Probabilities were calculated for all SNPs with $r^2 > 0.5$ to the lead SNP.

Because the calculation of thousands of permutations is computationally expensive and requires full genotype data, we sought to generalize the results of the permutation-based method in order to extend it to the analysis of autoimmune diseases for which Immunochip data were not available, or only the identity of the lead index SNPs was reported, such as from the GWAS catalogue. We developed a general model, where PICS was able to calculate $P(B^{\text{causal}}|A^{\text{lead}})$, where B is a SNP within a locus, and A is the lead SNP in the locus, by using LD relationships from the Immunochip where these were available, and from the 1000 Genomes Project otherwise. As the distribution of association signal at neutral SNPs in the locus approximates a normal distribution, given the lead SNP in the locus, we need to be able to estimate the mean expected association for a neutral SNP in LD with the lead SNP, and the standard deviation for that SNP.

The expected mean association signal for SNPs in the locus scales linearly with r^2 to the causal SNP in the locus. We derived an approximation for the standard deviation for each SNP in the locus based on the results of empiric testing. We picked 30,000 random SNPs from densely-mapped Immunochip loci, with half coming from the MS Immunochip data, and half coming from the IBD Immunochip data. For each SNP, we simulated 100 permutations with that SNP being the causal variant. SNPs selected had minor allele frequency above 0.05, and the odds ratio used varied from 1.1-fold to 2.0-fold. The number of cases and controls and total sample size were also allowed to randomly vary from 1–100% of the total number of samples in the original studies. These results indicated that the standard deviation for the association signal at a SNP in LD (with $r^2 > 0.5$) to a causal variant in the locus was approximately:

$$s = \sqrt{1 - r^k} \times \sqrt{\text{indexpval}} / 2$$

$$m = r^2 \times \text{indexpval}$$

where s is the standard deviation of the association signal at the SNP, m is the expected mean of the association signal at the SNP, indexpval is the $-\log_{10}(P \text{ value})$ of the causal SNP in the locus, r^2 is the square of the correlation coefficient (a measure of LD) between the SNP and the causal SNP in the locus, and k is an empiric constant that can be adjusted to fit the curve; in practice, we found that choosing k from a wide range of values between 6 and 8 had little measurable effect on the candidate causal SNPs selected, and we used a value of $k = 6.4$. The results of the 30,000 simulated iterations and the empiric curve fitted using the above equation is shown in Extended Data Fig. 3. To verify that our method was applicable to a wide range of case-control ratios and effect sizes, we performed six additional simulations, with the percentage of case samples fixed at 10%, 20%, and 50%, and the effect sizes of causal SNPs fixed at 1.2-fold, 1.5-fold, and 2.0-fold, which cover a broad range of parameters likely to be encountered in practical GWAS studies (Extended Data Fig. 3). We found that for all six scenarios, the relationship between r^2 to the causal SNP and standard deviation similarly followed the empirically fitted curve.

For each SNP in the locus, we used the estimated mean and standard deviation of the association signal at each neutral SNP in LD ($r^2 > 0.5$) to the lead SNP in the locus to calculate the probability of each SNP to be the causal variant relative to the lead SNP. We then normalized the probabilities so that the total of their probabilities summed to 1.

For diseases where summary SNP information was available, but the r^2 relationships between SNPs was unknown, the r^2 relationship was estimated based on the ratio between the association signal at the lead SNP versus the SNP in question. For diseases where only the lead SNP was known, r^2 values were drawn from the LD relationships from the MS Immunochip study if the SNP was from an Immunochip, or from the 1000 Genomes Project otherwise. 1000 Genomes European LD relationships were used for diseases, except for Kawasaki disease, for which 1000 Genomes East Asian LD relationships were used. For diseases that had both GWAS catalogue results and Immunochip results, we used Immunochip results whenever possible, and GWAS catalogue results in regions outside Immunochip dense-mapping coverage.

Multiple independent association signals. For the MS data, we were able to use full genotyping information to distinguish multiple independent signals. We used stepwise regression to condition away SNPs one at a time until no associations remain at the $P < 10^{-6}$ level, which is an effective method for separating independent signals, when LD between the independent causal variants is low. We then treated each independent signal separately for the purpose of using PICS to derive the likely causal variants.

Missing Immunochip data. For the minority of SNPs that were missing from the Immunochip, we used 1000 Genomes SNPs LD relations to the index SNP to estimate their probability of being the causal SNP. For the diseases with only Immunochip summary statistic data, we could not be certain of the LD relationships, and therefore we estimated the LD to the index SNP from the difference between the association at the lead SNP and the SNP in question, as these follow a linear relationship. For the diseases that only had Immunochip index SNP data, we used Immunochip LD relationships where available from the MS data, and 1000 Genomes SNPs LD relations to the index SNP where these were not available.

Distance between GWAS catalogue SNPs and lead SNPs. For Immunochip regions that were previously studied by non-Immunochip studies, we examined the performance of prior non-fine-mapped studies at correctly determining the lead SNP. GWAS catalogue SNPs within 200 kb of Immunochip regions were considered, and the LD and genomic distance between the catalogue SNP and any Immunochip lead SNPs for that disease in the Immunochip region were measured and reported in the histograms in Fig. 1d and Extended Data Fig. 5. PICS was also used to calculate the probability of GWAS catalogue SNPs to be causal variants; the probability was 5.5% on average.

Number of candidate causal SNPs per GWAS signal. For each GWAS signal, we obtained a set of candidate causal SNPs, each with a probability of being the causal variant. For each signal, we asked what was the minimum number of candidate causal SNPs required to cover at least 75% of the probability (Fig. 1e).

Distribution of GWAS signals in functional genomic elements: signal to background. For downstream analyses, we considered the set of 4,905 candidate causal SNPs (the cutoff was probability > 0.0275). We performed 1,000 iterations, picking 4,905 minor-allele-frequency-matched random SNPs from the same loci (from genomic regions within 50 kb of the candidate causal SNPs and excluding the actual causal SNPs). It was necessary to match for minor-allele-frequency because lower MAF SNPs are far more likely to be coding variants. Furthermore, it was necessary to match for locus, because GWAS SNPs are greatly enriched at gene bodies, and using a background of random 1,000 genome SNPs for comparison results in massive non-specific enrichment of all functional elements. Because we are comparing the candidate causal SNPs to a background set of control SNPs from the same regions, the observed enrichments at functional elements strongly argues that PICS effectively predicts causal variants within the loci. For each functional category (missense, nonsense, and frameshift were merged), we calculated the number of actual candidate causal SNPs above mean background (mean of 1,000 random iterations), divided by the total number of GWAS signals represented (635), and used these results to populate the pie chart indicating the approximate percentage of GWAS signals that can be attributed to each assessed functional category (Fig. 6).

Analysis of ex vivo stimulation-dependent enhancers. We searched for motifs enriched in cell type-specific enhancers in the five stimulated T-cell subsets (PMA/ionomycin stimulated TH_{stim} and $TH17$ T cells, anti-CD3/CD28 stimulated $TH0$, $TH1$, and $TH2$ T cells) compared to enhancers in naive T cells, using the motif finding program HOMER (<http://homer.salk.edu/homer/>)⁶³. AP-1 was the most strongly enriched motif in enhancers that gained H3K27ac in the stimulated T cells (Fig. 2), whereas this enrichment was absent when comparing naive T cells with memory or regulatory T cells. Additional motifs that were enriched in the stimulation-dependent enhancers included NFAT for the PMA/ionomycin stimulation conditions and STAT for the anti-CD3/CD28 stimulation conditions.

Enhancer signal-to-noise analysis. We focused on 14 immune cell types (8 $CD4^{+}$ T-cell subsets, 2 $CD8^{+}$ T-cell subsets, $CD14^{+}$ monocytes, and 3 B-cell subsets) and 19 representative non-immune cell/tissue types from the Roadmap Epigenomics project. Enhancers were broken up into 1 kb segments and immune specific enhancers were identified based on the following criteria: (1) number of normalized mean H3K27ac ChIP-seq extended reads/base > 4 , and (2) mean H3K27ac in the top fifteenth percentile when comparing immune cells to non-immune cells/tissues. We measured the percentage of PICS SNPs (with different probability cutoffs) that either map to an immune enhancer or cause an amino acid coding change (Fig. 2). We next considered the 4,300 candidate causal SNPs that were not associated with protein-coding changes, and compared them against 1,000 iterations of frequency and locus matched controls (picked from genomic regions within 50 kb of the candidate causal SNPs and excluding the actual candidate causal SNPs; see discussion of background calculations above). Enhancers were enriched approximately 2:1 above background. We also measured the signal-to-background ratio for GWAS

signals that had been attributed to coding variants; these produced a much lower signal to background ratio for immune enhancers, as would be anticipated by the fact that most of these are acting on coding regions rather than enhancers (Extended Data Fig. 7). The mean signal above background was shown in a pie chart (Fig. 6).

Comparison to other methods for determining candidate causal variants. We compared the efficacy of PICS versus previously published methods used to determine candidate causal variants (Fig. 2d, e). We first considered studies that had used cutoffs of $r^2 = 1.0$ and $r^2 > 0.8$ to determine likely causal SNPs. Because prior studies had not made use of dense genotyping data, we used only the GWAS catalogue results for this comparison, and applied PICS, and the two r^2 -cutoff criteria. In practice these were much more stringent than prior analyses, because we limited the GWAS catalogue studies to those that produced 6 or more genome-wide significant hits, thereby pruning underpowered studies. We also required a significance of $P < 10^{-6}$ for index SNPs, and merged index SNPs at the same locus to use the strongest and most accurate lead SNP. We found that PICS autoimmunity SNPs were much more likely to map to immune enhancers than SNPs identified by the other statistics. In addition, when the PICS SNPs which overlapped the $r^2 > 0.8$ and $r^2 > 1.0$ sets were removed, the remaining SNPs did not show any enrichment above background. In contrast, the candidate causal SNPs identified by PICS, but missed by both of the other methodologies, were significantly enriched for immune enhancers. Background was calculated based on random SNPs drawn from the same loci (within 50 kb, frequency-matched controls) as the candidate causal SNPs.

We also compared PICS with a recently reported Bayesian approach⁶⁴, using a recently published study of MS²⁰ that employed this methodology to call candidate causal SNPs. Because this published method required full genotypes to be available, this comparison was limited to only the MS dataset. Both PICS and the published method are Bayesian approaches, where each SNP within the locus is given equal prior consideration to be the causal variant, and the algorithm then weighs each SNP based on the likelihood of each model given the data. However, the PICS method provides two advantages. First, the probabilities assigned to each SNP by PICS are determined empirically using permutation, rather than using a theoretical estimate for the weight of each SNP. Second, PICS can be generalized to all GWAS data with publicly available summary statistic data and does not rely on genotype data.

For the same MS Immunochip data set, PICS called 434 candidate causal SNPs, whereas the prior method called 4,070 candidate causal SNPs; 177 SNPs were shared between the two analyses. Of the 434 PICS candidate causal SNPs, 26.5% overlapped immune enhancers, whereas 9.5% of the SNPs from the other method overlapped immune enhancers; the background rate of random SNPs from the same loci overlapping immune enhancers was 8% (Extended Data Fig. 4). Because the method⁶⁴ is clearly less stringent than PICS, we also tried using a high confidence set of SNPs derived by that method, by selecting the top SNPs such that their average probability of being a causal variant was 10% (the same cutoff used for the PICS SNPs). There were 165 SNPs in this high confidence set, compared to 434 for PICS, with an overlap of 65 SNPs. 20.3% of the candidate causal SNPs in the high confidence set⁶⁴ overlapped immune enhancers. Although anecdotal, these results suggest that PICS performs at least as well as the prior method.

Tissue-specificity of diseases. We used PICS fine mapping to determine the set of candidate causal SNPs for each of 39 different diseases, and examined whether they were enriched within the enhancers most specific to each cell type (defined as being in the top fifteenth percentile of H3K27ac signal compared to other cell types, and with > 1 normalized mean H3K27ac ChIP-seq extended reads/base). To compare enhancer regions across different cell types, we first subdivided regions of the genome that were marked as enhancers into enhancer segments ~ 1 kb in size. Next, H3K27ac read density at each enhancer segment in the genome was compared across all 33 cell types to determine the cell types in the top fifteenth percentile (H3K27ac signal was quantile normalized across the cell types before comparison). The heatmap (Fig. 3) depicts P values for the enrichment of PICS SNPs for each disease in H3K27ac elements for each cell type, as calculated by the chi square test. For this comparative analysis, enrichment of PICS SNPs was measured against a background of all common 1000 Genomes SNPs. We used this approach because the goal was to highlight cell-type-specificity of the diseases, which would have been normalized out by the rigorous locus controls used above, and given that the specificity of PICS SNPs for enhancers within the loci was already established. We also mapped the expression patterns of genes with PICS candidate causal coding SNPs associated with Crohn's disease, MS and rheumatoid arthritis (Extended Data Fig. 8).

Super-enhancer enrichment. The full set of loci called as super-enhancers¹⁸ in $CD4^{+}$ T-cell subsets (T_{naive} , T_{mem} , $TH17$, TH_{stim}) were merged and identified as $CD4^{+}$ T-cell super-enhancer regions. These regions often contain clusters of discrete enhancers marked with H3K27ac, separated by non-acetylated regions. We assessed if PICS SNPs mapping to super-enhancers were more likely to occur in H3K27ac-marked enhancer regions than in intervening regions. Within $CD4^{+}$ T-cell super-enhancer regions, we compared overlap of PICS candidate causal SNPs with $CD4^{+}$ T-cell H3K27ac regions, compared to frequency-matched background SNPs drawn from

these same regions (Extended Data Fig. 7). H3K27ac intervals in CD4⁺ T-cell super-enhancers were called based on being in the top fifteenth percentile in mean H3K27ac in T cells compared to the other 25 cell types. In addition, we assessed overlap between PICS SNPs and H3K27ac elements preferential to either stimulated or unstimulated CD4⁺ T cells. Stimulated CD4⁺ T-cell elements were defined as those with a mean increase of > 25% in H3K27ac in the (average of) TH17, TH1, TH2, TH0, TH1, TH2 cells, compared to the T_{naive}, T_{mem}, T_{reg}; the remainder of the CD4⁺ T-cell set were defined as unstimulated elements.

Figure 4 shows that some sub-elements within *IL2RA* super-enhancer locus appear bound by T-cell master regulators based on published ChIP-seq data, including FOXP3 in T_{regs}, T-bet in TH1 cells, and GATA3 in TH2 cells^{65,66}.

Non-coding RNA analysis. We next examined the set of disease-associated enhancers, that is, immune enhancers containing PICS autoimmunity SNPs, and their association with non-coding RNAs. Non-coding RNA transcripts were called based on a RNA-seq read density of 0.5 genome-normalized reads per bp over a window size of at least 2 kb, excluding RNA transcripts overlapping annotated exons or gene bodies. We found that enhancers containing PICS autoimmunity SNPs were enriched for non-coding transcript production, primarily consistent with unspliced enhancer-associated RNAs. Candidate causal SNPs were enriched 1.6-fold within T-cell enhancers that transcribed non-coding RNAs, compared to T-cell enhancers overall ($P < 0.01$).

H3K27ac and DNase profiles. We measured H3K27ac profiles and DNase hypersensitivity profiles in a 12 kb window centred around candidate causal SNPs, taking the average signal for the 14 immune cell types for which H3K27ac was available, and immune cell types from ENCODE²⁶ for which DNase was available (CD14⁺, GM12878, CD20⁺, TH17, TH1, TH2). Average normalized reads for H3K27ac and DNase centred at PICS SNPs are displayed in Fig. 5a.

Transcription factor ChIP-seq binding site analysis. We compared the enrichment of PICS autoimmunity SNPs at transcription factor binding sites identified by ENCODE ChIP-seq⁶⁷, relative to random SNPs drawn from the same loci (50 kb window around the candidate causal SNPs, frequency matched). We show the results for the 31 transcription factors whose binding sites are most significantly enriched for PICS SNPs (Fig. 5b).

Motif creation / disruption analysis. We downloaded consensus motifs from SELEX²⁸ and Xie *et al.*²⁹ (represented as degenerate nucleotide codes). We used the 853 highest probability non-coding PICS SNPs (mean probability = 0.30, cutoff > 0.1187), representing 403 different GWAS signals. For each candidate causal SNP, we examined whether it created or disrupted a known motif from SELEX or Xie *et al.*²⁹ For comparison, we ran 1,000 iterations using frequency-matched random SNPs drawn from the same loci (within 50 kb of the PICS SNPs). We found several known motifs (Extended Data Fig. 9) to be significantly enriched, including AP1, ETS, NF-KB, SOX, PITX, as well as several unknown conserved motifs (Extended Data Fig. 9). Subtracting the number of motifs found to be disrupted against that expected by background, and dividing by the total number of GWAS signals, we estimate that approximately 11% of non-coding GWAS hits can be attributed to direct disruption or creation of transcription factor binding motifs.

Neighbouring motif analysis. We compared the sequence within 100 nt of high-likelihood PICS SNPs (cutoff > 0.1187) against random flanking sequence (10 kb away on either side from the causal SNPs) and looked for enriched motifs using HOMER (<http://homer.salk.edu/homer/>)⁶³. We found significant enrichments for NF-KB, RUNX, AP1, ELF1, and PU1 (Extended Data Fig. 9). Interestingly, there was a palindromic unknown motif TGGCWNNNWGCCA ($P < 10^{-4}$) previously defined by phylogenetic conservation that was significant both in this method and in the motif disruption/creation analysis. This motif resembles the consensus motif for Nuclear Factor I (NFI) transcription factors, suggesting a role for at least one member of this transcription factor family in autoimmunity.

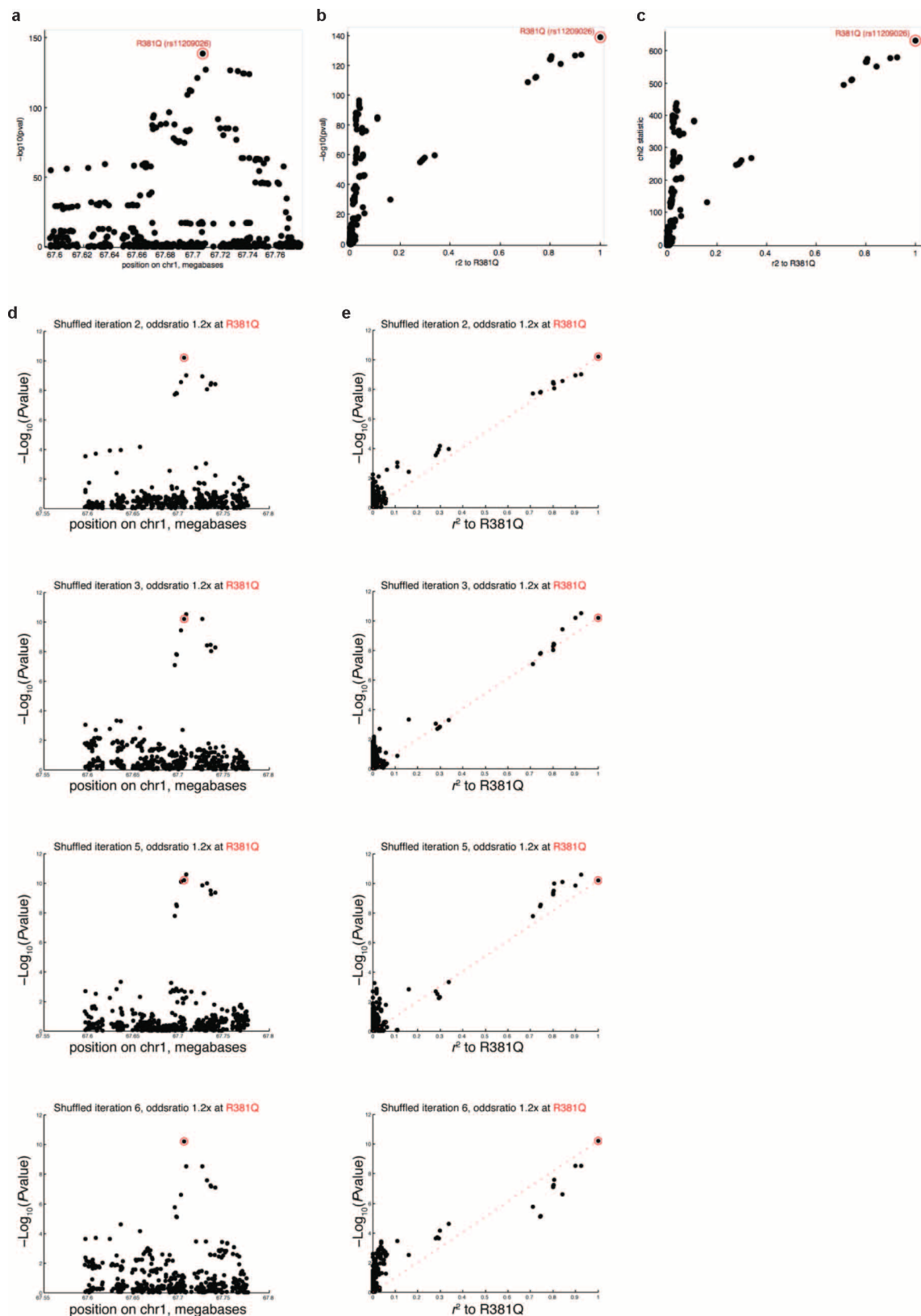
Expression quantitative trait loci (eQTL) analysis. We used PICS to predict causal SNPs from a peripheral blood eQTL data set with 1000 Genomes summary statistic data available for all *cis*-eQTLs. We required a gene to have a *cis*-eQTL with a P value < 10^{-6} for this analysis, giving us 4,136 genes. For each gene we applied PICS. We considered a autoimmunity GWAS hit to score as an eQTL if any autoimmunity PICS SNP in the locus coincided with an eQTL PICS SNP with average probability > 0.01%. We found that 11.6% (74/636) of autoimmunity GWAS hits were also eQTL hits. In addition, 18.5% (15/81) of coding GWAS hits also showed eQTL effects, suggesting that they may actually operate at the transcriptional level, in addition to any coding effects they may have.

To quantify overlap of candidate causal eQTL SNPs with functional elements, we compared PICS eQTL SNPs against frequency-matched background SNPs drawn from the same loci (within 50 kb) in 1,000 iterations. These comparisons are shown in signal-to-background bar graphs for both coding/transcript-related functional elements and for enhancers and promoters (Extended Data Fig. 10). The signal above mean background was calculated for each functional category, and these results were

compared against the results for autoimmunity GWAS hits in the pie charts shown in Fig. 6a.

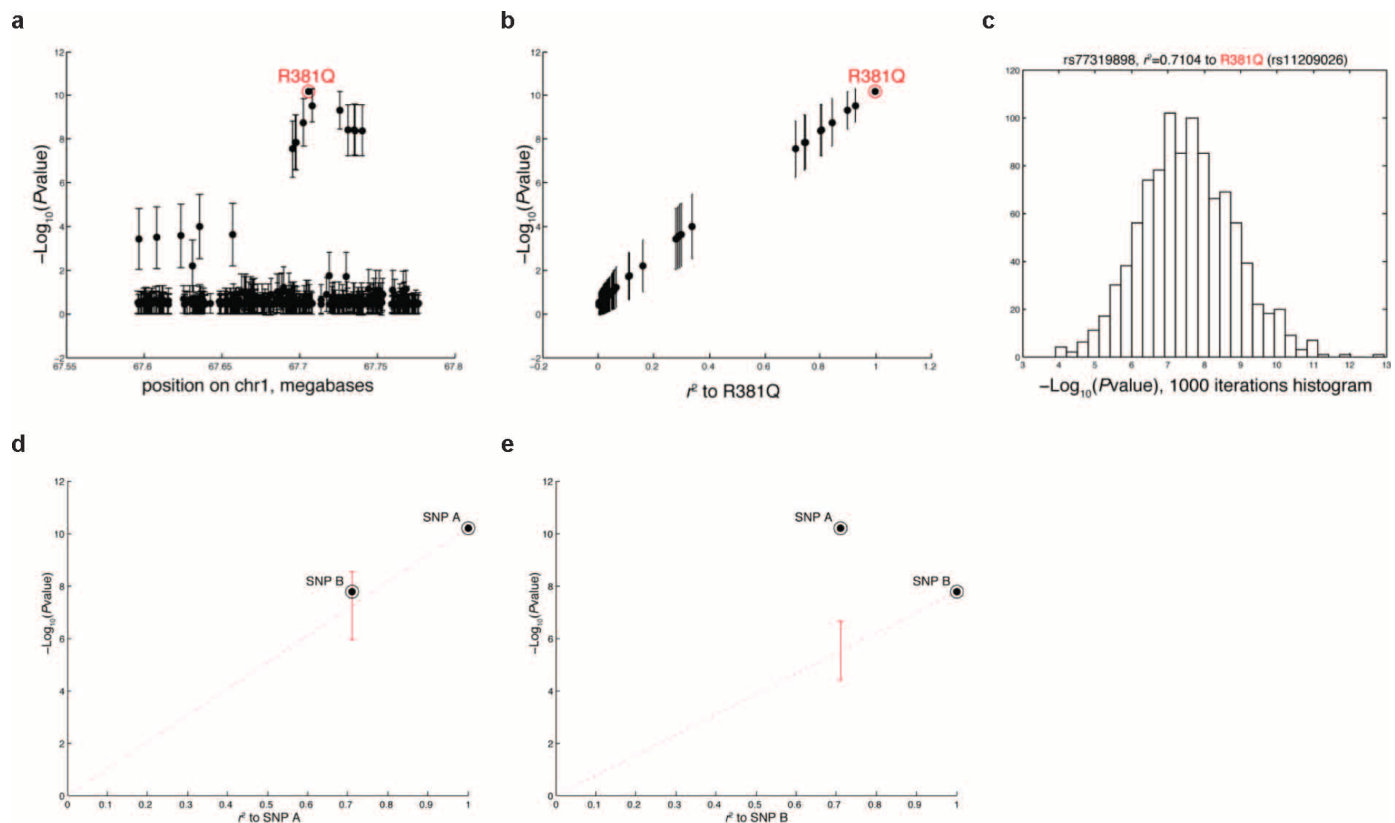
We further examined whether the magnitudes of disease-associated eQTLs differed, compared to the space of all eQTLs (Extended Data Fig. 10). Disease-associated variants had modestly larger effects on gene expression ($P < 10^{-6}$ by rank-sum test), but did not necessarily correspond to the strongest eQTLs.

35. Brucklacher-Waldert, V. *et al.* Phenotypical characterization of human Th17 cells unambiguously identified by surface IL-17A expression. *J. Immunol.* **183**, 5494–5501 (2009).
36. Johnston, A., Sigurdardottir, S. L. & Ryon, J. J. Isolation of mononuclear cells from tonsillar tissue. *Current Protoc. Immunol.* <http://dx.doi.org/10.1002/0471142735.im0708s86> (2009).
37. Caron, G., Le Gallou, S., Lamy, T., Tarte, K. & Fest, T. CXCR4 expression functionally discriminates centroblasts versus centrocytes within human germinal center B cells. *J. Immunol.* **182**, 7595–7602 (2009).
38. Zhu, J. *et al.* Genome-wide chromatin state transitions associated with developmental and environmental cues. *Cell* **152**, 642–654 (2013).
39. Gifford, C. A. *et al.* Transcriptional and epigenetic dynamics during specification of human embryonic stem cells. *Cell* **153**, 1149–1163 (2013).
40. Engreitz, J. M. *et al.* The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. *Science* **341**, 1237973 (2013).
41. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
42. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357–359 (2012).
43. Béguelin, W. *et al.* EZH2 is required for germinal center formation and somatic EZH2 mutations promote lymphoid transformation. *Cancer Cell* **23**, 677–692 (2013).
44. Beyer, M. *et al.* High-resolution transcriptome of human macrophages. *PLoS ONE* **7**, e45466 (2012).
45. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001–D1006 (2014).
46. Hindorf, L. A. *et al.* A catalog of published genome-wide association studies. (<http://genome.gov/gwastudies> accessed July 2013).
47. Trynka, G. *et al.* Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nature Genet.* **43**, 1193–1201 (2011).
48. Cooper, J. D. *et al.* Seven newly identified loci for autoimmune thyroid disease. *Hum. Mol. Genet.* **21**, 5202–5208 (2012).
49. Liu, J. Z. *et al.* Dense fine-mapping study identifies new susceptibility loci for primary biliary cirrhosis. *Nature Genet.* **44**, 1137–1141 (2012).
50. Eyre, S. *et al.* High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. *Nature Genet.* **44**, 1336–1340 (2012).
51. International Genetics of Ankylosing Spondylitis Consortium. Identification of multiple risk variants for ankylosing spondylitis through high-density genotyping of immune-related loci. *Nature Genet.* **45**, 730–738 (2013).
52. Ellinghaus, D. *et al.* High-density genotyping study identifies four new susceptibility loci for atopic dermatitis. *Nature Genet.* **45**, 808–812 (2013).
53. Liu, J. Z. *et al.* Dense genotyping of immune-related disease regions identifies nine new risk loci for primary sclerosing cholangitis. *Nature Genet.* **45**, 670–675 (2013).
54. Hinks, A. *et al.* Dense genotyping of immune-related disease regions identifies 14 new susceptibility loci for juvenile idiopathic arthritis. *Nature Genet.* **45**, 664–669 (2013).
55. Tsoi, L. C. *et al.* Identification of 15 new psoriasis susceptibility loci highlights the role of innate immunity. *Nature Genet.* **44**, 1341–1348 (2012).
56. Pidasheva, S. *et al.* Functional studies on the IBD susceptibility gene IL23R implicate reduced receptor function in the protective genetic variant R381Q. *PLoS ONE* **6**, e25038 (2011).
57. Duerr, R. H. *et al.* A genome-wide association study identifies *IL23R* as an inflammatory bowel disease gene. *Science* **314**, 1461–1463 (2006).
58. Armitage, P. Tests for linear trends in proportions and frequencies. *Biometrics* **11**, 375–386 (1955).
59. Franke, A. *et al.* Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature Genet.* **42**, 1118–1125 (2010).
60. The UK IBD Genetics Consortium and The Wellcome Trust Case Consortium 2. Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the *HNF4A* region. *Nature Genet.* **41**, 1330–1334 (2009).
61. Barrett, J. C. *et al.* Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nature Genet.* **40**, 955–962 (2008).
62. Anderson, C. A. *et al.* Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nature Genet.* **43**, 246–252 (2011).
63. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
64. Wellcome Trust Case Control Consortium. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nature Genet.* **44**, 1294–1301 (2012).
65. Birzele, F. *et al.* Next-generation insights into regulatory T cells: expression profiling and FoxP3 occupancy in human. *Nucleic Acids Res.* **39**, 7946–7960 (2011).
66. Kanhere, A. *et al.* T-bet and GATA3 orchestrate Th1 and Th2 differentiation through lineage-specific targeting of distal regulatory elements. *Nature Commun.* **3**, 1268 (2012).
67. Gerstein, M. B. *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91–100 (2012).



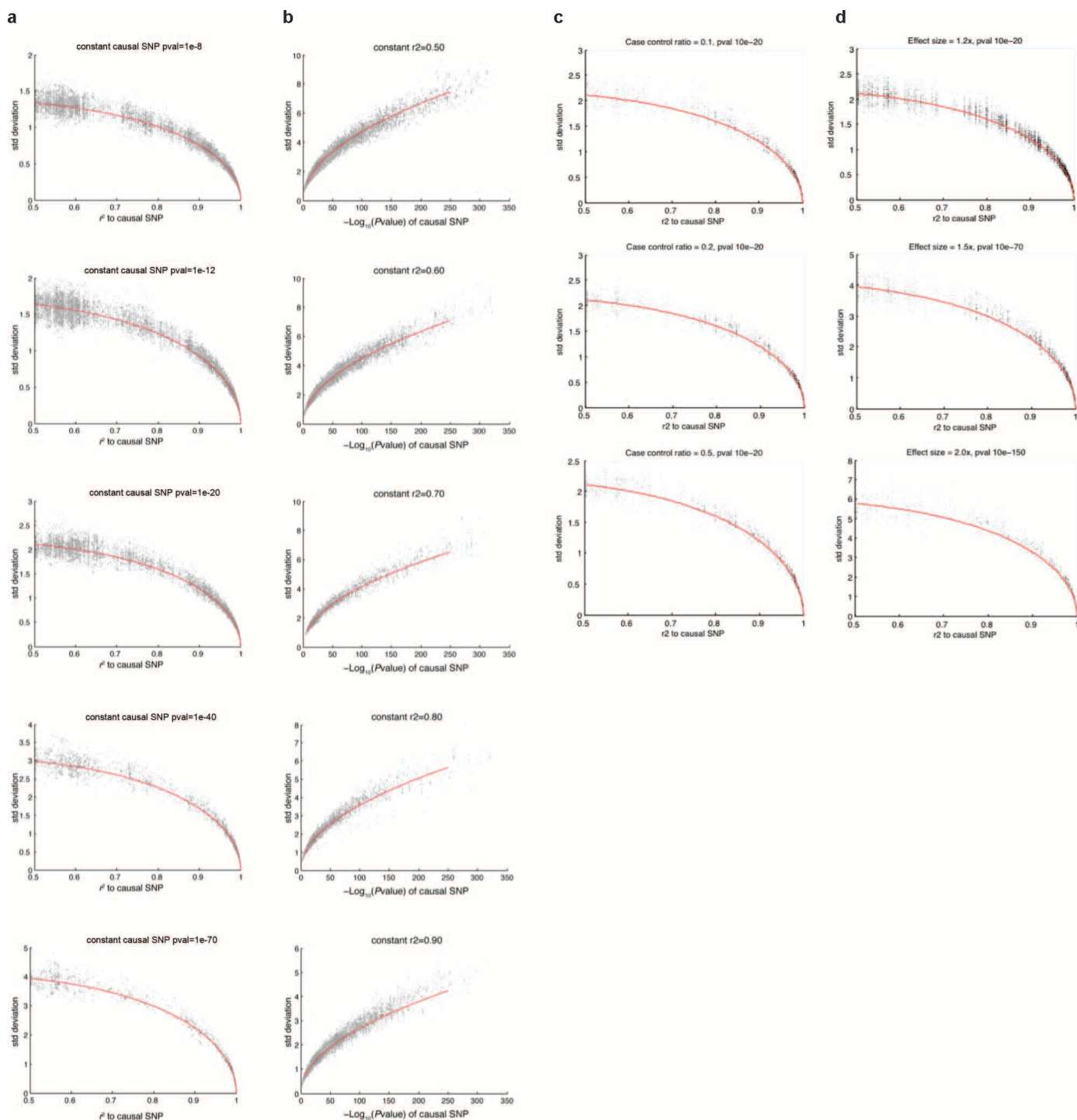
Extended Data Figure 1 | GWAS result for IBD Immunochip data at *IL23R* locus. **a**, Each of the 500 SNPs in the *IL23R* densely genotyped locus is plotted according to its association signal and position along the chromosome. The R381Q missense variant is circled in red. **b**, Each of the 500 SNPs in the *IL23R* densely genotyped locus is plotted according to its association signal and r^2 linkage to R381Q. **c**, Same as **b**, but showing the association signal on the y axis in χ^2 units. Over the range of values typically encountered in GWAS analyses, χ^2 units and log P value are asymptotically linear. **d**, Simulated permutation

analysis of signal at *IL23R* locus. The 1.2-fold odds ratio signal was simulated at the R381Q SNP by fixing the association signal at R381Q, but permuting cases and controls such that all other SNPs are neutral and vary only with statistical noise. Four representative results from the simulations are shown, with the panels on the left showing the association signal in genomic space, and the panels on the right (**e**) showing the association signal for each SNP in relation to r^2 . Actual data is shown in **a–c**, simulated permutation is shown in **d, e**.



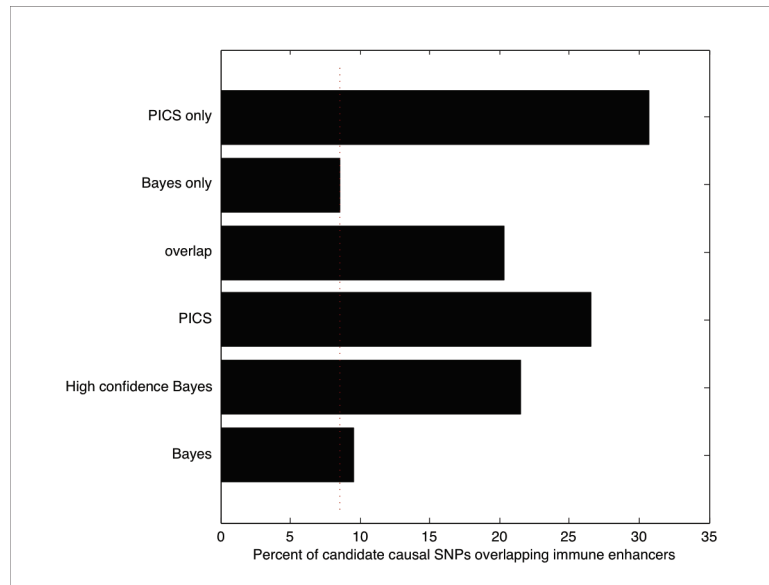
Extended Data Figure 2 | Calculating the relative likelihood of being the causal SNP from standard deviation in association signal. **a, b,** For each SNP in the *IL23R* locus, the mean association signal and the standard deviation, calculated across 1,000 permutations (using a 1.2-fold odds ratio at the R381Q SNP), are shown in genomic space (**a**) and in terms of each SNP's r^2 linkage disequilibrium to the causal R381Q variant (**b**). **c,** The distribution of association signals at rs77319898 ($r^2 = 0.71$ to the causal variant) for 1,000 permutations is shown. The distribution of association signal values at each SNP approximated a normal distribution. **d,** PICS analysis of a two SNP case to determine the relative likelihood of each to explain the pattern of association at

the locus. The SNPs represented here are R381Q (SNP A) and rs77319898 (SNP B), which has an $r^2 = 0.71$ to R381Q. The signal at SNP B is well-explained by LD to SNP A, in a model where SNP A is treated as the putative causal variant. The error bars indicate the standard deviation in the association signal expected for SNP B, under the assumption that SNP A is causal. **e,** The signal at SNP A is poorly explained by LD to SNP B, in a model where SNP B is treated as the putative causal variant. The error bars indicate the standard deviation in the association signal expected for SNP A, under the assumption that SNP B is causal.



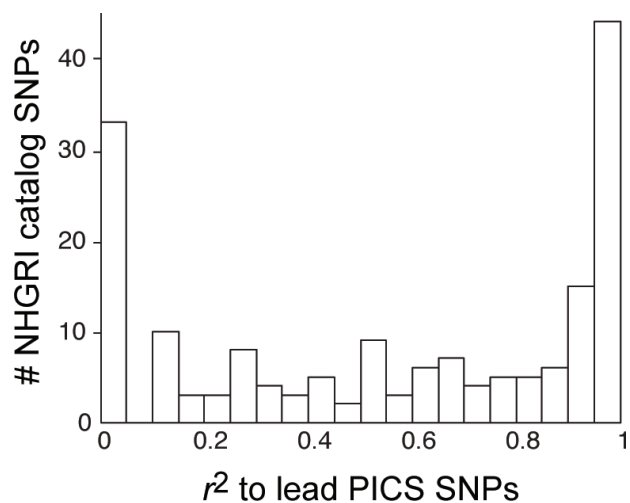
Extended Data Figure 3 | Simulated permutations and empiric curve fitting for 30,000 GWAS signals at Immunochip loci. **a**, We simulated 30,000 causal SNPs in densely mapped Immunochip regions. Plot shows the relationship between standard deviation in the association signal of neutral SNPs and their r^2 to the causal SNP (neutral SNPs within $r^2 > 0.5$ of the simulated causal variant are shown). The red line indicates the expected values derived from the empiric equation for the standard deviation of the association signal at neutral SNPs in LD with the causal SNP. **b**, Plot shows the relationship between standard deviation in the association signals of neutral SNPs and the association signal of the causal SNP. Each panel represents the set of neutral SNPs with the indicated r^2 to the causal variant. **c**, Simulated permutations over

a range of case-control ratios. We plotted the relationship between standard deviation at neutral SNPs and their r^2 to the causal SNP. Plots are shown for three series of simulations, with the percentage of cases fixed at 10%, 20%, and 50% of the total sample size, and a causal SNP P value of 10^{-20} . Red line indicates the expected values derived from the empiric equation for the standard deviation of the association signal at neutral SNPs in LD with the causal SNP in the locus. **d**, Simulated permutations over a range of effect sizes. Plots are shown for three series of simulations, with the effect size fixed at 1.2-fold, 1.5-fold, and 2.0-fold, and the corresponding lead SNP P values fixed at 10^{-20} , 10^{-70} , and 10^{-150} , respectively.

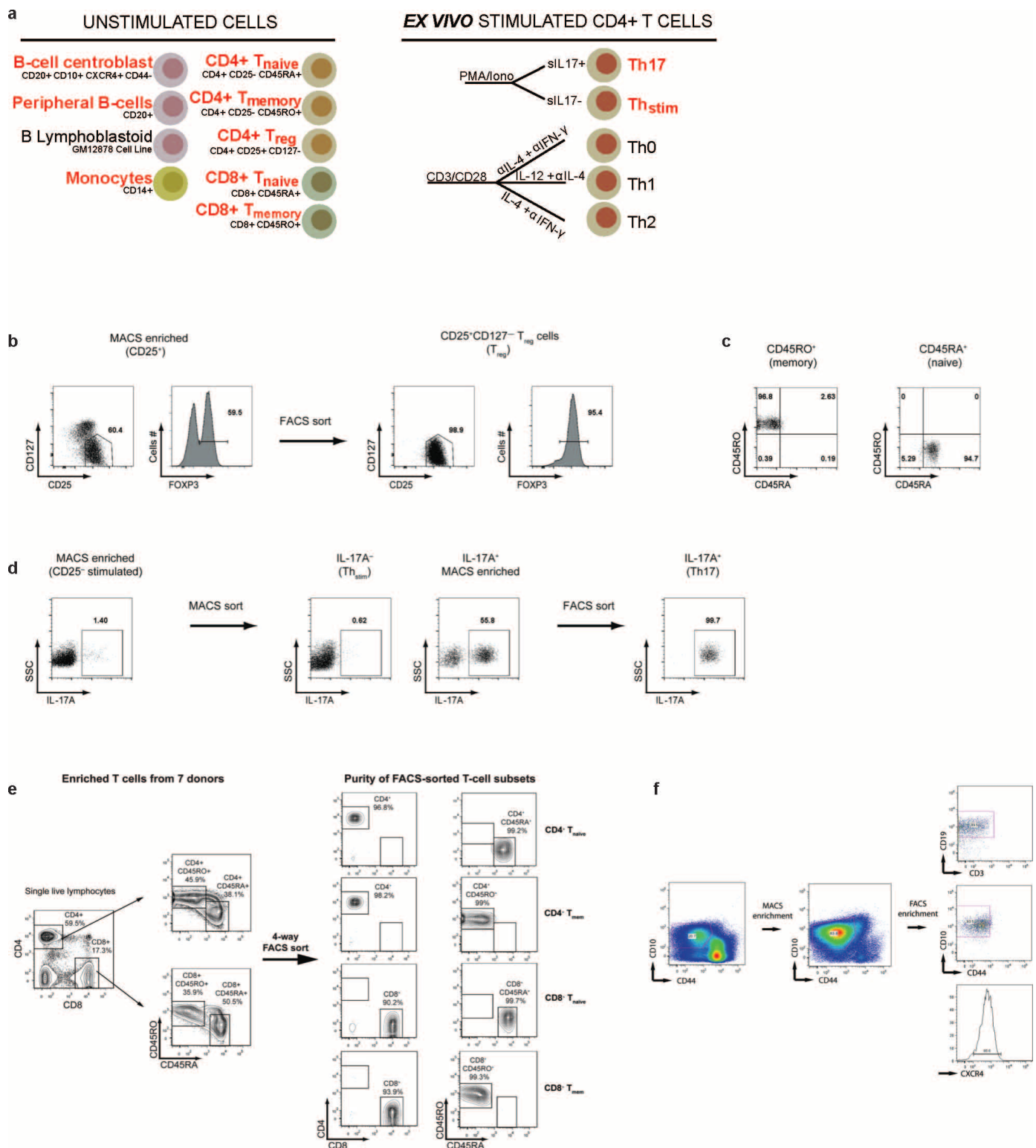


Extended Data Figure 4 | Comparison of PICS with prior Bayesian fine-mapping method. Bar graph shows the percentage of MS SNPs overlapping immune enhancers using different algorithms for calling candidate causal SNPs. The dotted line indicates the background rate at which random 1000 Genomes Project SNPs drawn from the same loci intersect immune

enhancers (~8%). The categories shown are (from top to bottom): 257 SNPs called only by PICS, 3,812 SNPs called only by the Bayesian method, 177 SNPs called by both PICS and the Bayesian method, all 434 SNPs called by PICS, 165 called by the Bayesian using a cutoff that only includes the highest confidence SNPs, and all 4,070 SNPs called by Bayesian method.



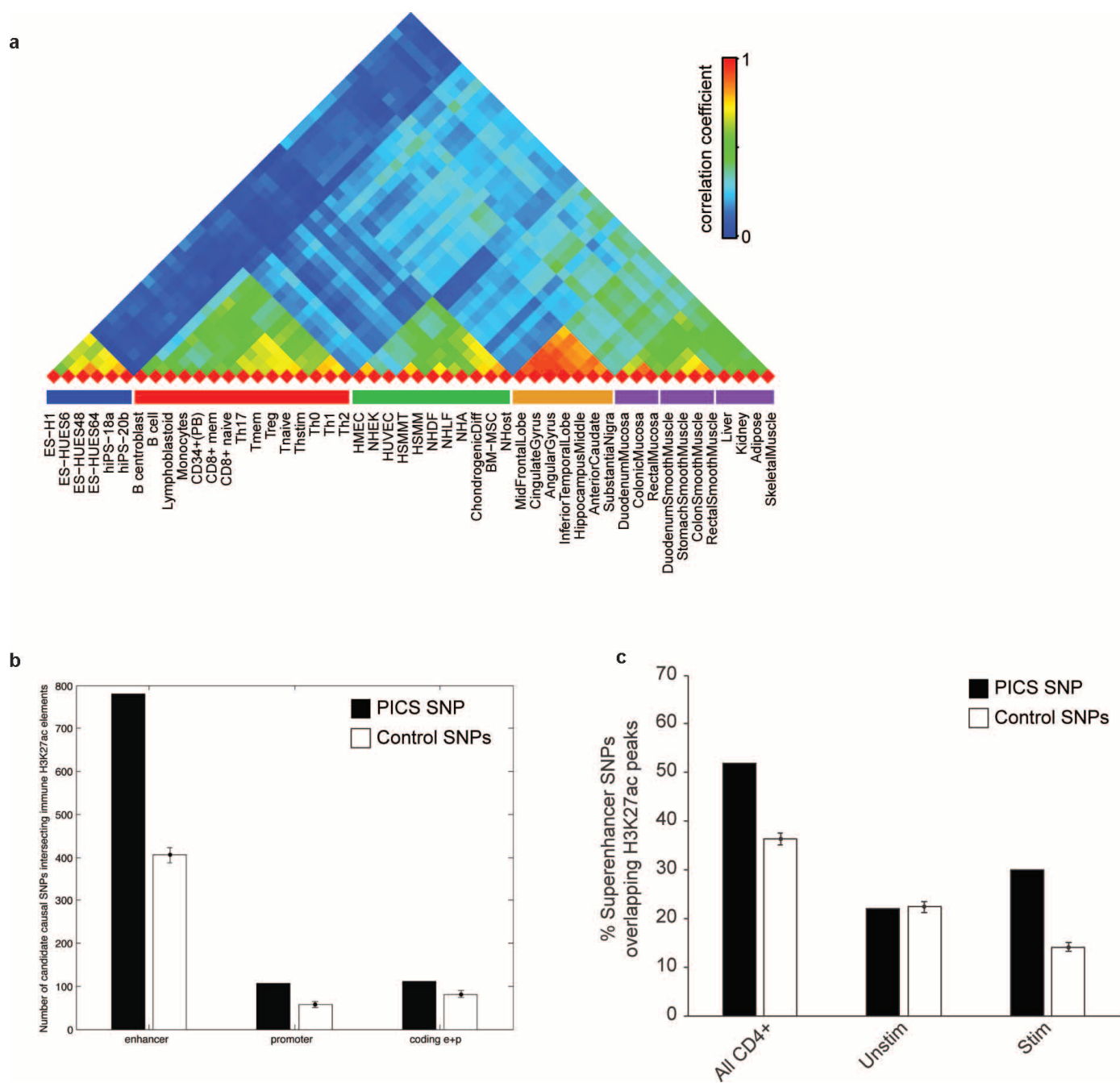
Extended Data Figure 5 | LD distance between PICS lead SNPs and GWAS catalogue index SNPs. Histogram indicates LD distance (in r^2) between PICS fine-mapped Immunochip lead SNPs and previously reported GWAS catalogue index SNPs from the same loci.



Extended Data Figure 6 | Purification of human immune cell subsets.

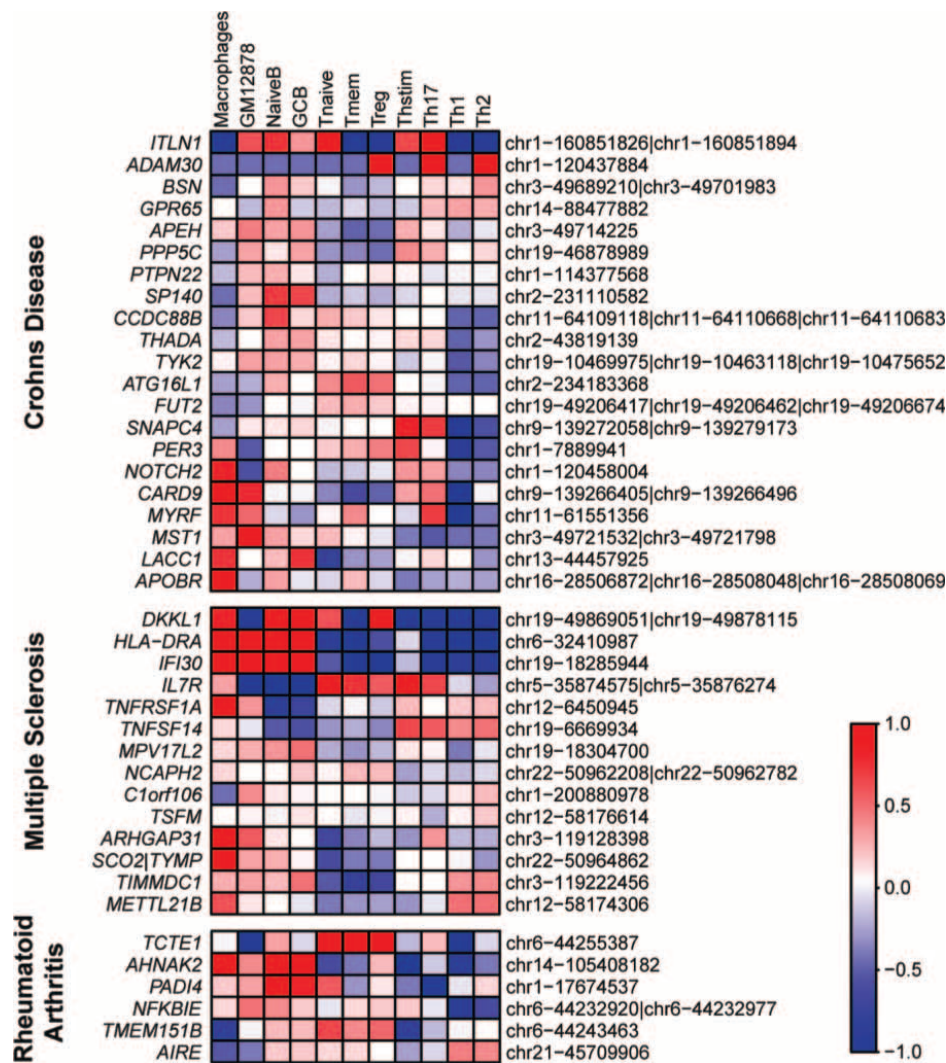
a, Immune populations subjected to epigenomic profiling in this study (red labels) or prior publications. **b**, CD4⁺ cells were enriched based on CD25 expression (MACS) and subsequently sorted based on CD25^{hi}CD127^{lo} to isolate T_{reg} cells; confirmed with FOXP3 intracellular staining. **c**, CD4⁺CD25⁺ cells were sorted to isolate T_{mem} (CD45RO⁺CD45RA⁻) and T_{naive} (CD45RO⁻CD45RA⁺) cells. **d**, CD4⁺CD25⁺ cells were PMA/ionomycin

stimulated and separated based on IL17 surface expression (MACS and FACS) to isolate Th17 cells (IL17⁺) and Th_{stim} cells (IL17⁻). **e**, Naive (CD45RA⁺CD45RO⁻) and memory (CD45RA⁻CD45RO⁺) CD8⁺ T cells were isolated using a BD FACSaria 4-way cell sorter. Results are shown from one of two large-scale sorts. **f**, Mononuclear cells were isolated from paediatric tonsils. Following CD10 enrichment (MACS), B centroblasts (CD19⁺CD10⁺CXCR4⁺CD44⁻CD3⁻) were purified by FACS.



Extended Data Figure 7 | PICS SNPs localize to immune enhancers and stimulus-dependent H3K27ac peaks in super-enhancers. **a**, Correlation matrix of 56 cell types, clustered by similarity of H3K27ac profiles (high = red, low = blue). **b**, Enrichment of non-coding autoimmune disease candidate causal SNPs within immune enhancers and promoters compared to background. The background expectation is based on frequency-matched control SNPs drawn from within 50 kb of the candidate causal SNPs. Candidate causal SNPs that produced coding changes or were in LD with a coding variant

(paired bars on the right) showed a smaller degree of enrichment in immune enhancers and promoters compared to background. **c**, Overlap of PICS SNPs with H3K27ac peaks within T-cell super-enhancers. Bar plot shows overlap of PICS SNPs with H3K27ac peaks in super-enhancers in CD4⁺ T-cells, compared to random SNPs drawn from within the same super-enhancers (all CD4⁺; left bar graph). Adjacent bars show overlap to H3K27ac peaks within CD4⁺ T-cell super-enhancers that do (Stim) or do not (Unstim) increase their acetylation upon stimulation.



Extended Data Figure 8 | Expression pattern of genes with PICS autoimmunity coding SNPs. Heatmap shows the relative expression levels of genes with coding SNPs associated with Crohn's disease, multiple sclerosis, and rheumatoid arthritis.

a

Known motifs created or disrupted by candidate causal SNPs

Motif	Observed	Expected	Pvalue <	Annotation
RRACAATG	8	1.6	10^{-3}	SOX
CAGGAARY	5	.82	.01	ETS/ELF1
TGANTCA	8	2.63	.03	AP-1
CCACTTRA	2	.12	.05	NKX2-3
GCTKASTCA	2	.12	.02	MAFK
TTAATCC	2	.24	.05	PITX1
GGGAWWTCC	2	.28	.05	NFKB

b

Additional motifs created or disrupted by candidate causal SNPs

Motif	Observed	Expected	Pvalue <	Annotation
KMCATNNWGA	7	.45	10^{-5}	XIE116
TGGNNNNNNKCCAR	4	.65	.01	XIE27
WYAAANNRNNNGCG	2	.12	.02	XIE126
CCNNNNNNNAAGWT	3	.41	.02	XIE158
ATTTCAW	6	1.97	.03	XIE174
CTGRNNNTTGW	3	.61	.04	XIE152

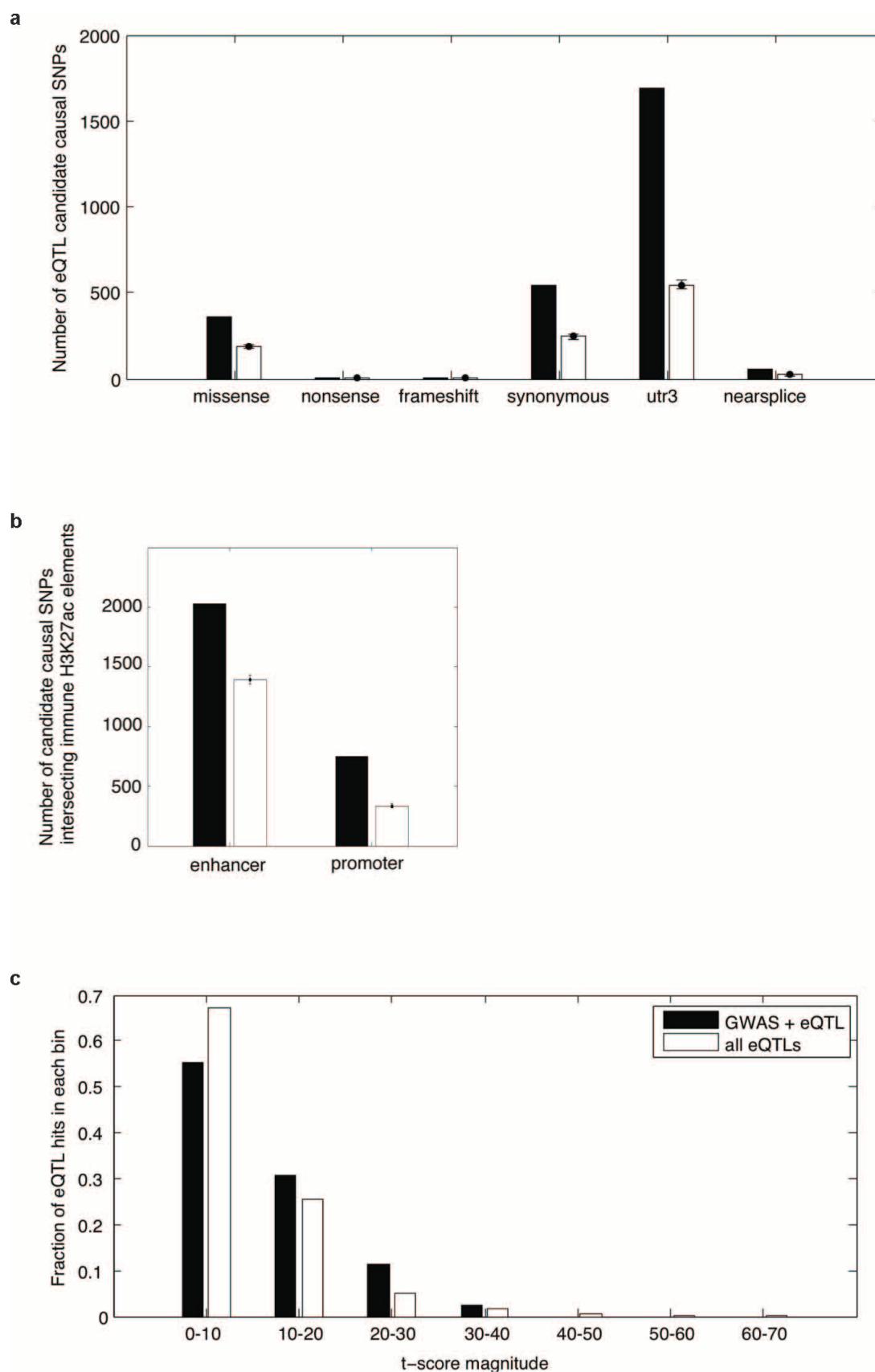
c

Known motifs enriched within 100bp of candidate causal SNPs

Motif	Observed	Expected	Pvalue	Annotation
GGAAATTCCC	7.78%	4.53%	$p < 10^{-4}$	NFKB
WAACCACAR	9.04%	5.82%	$p < 10^{-4}$	RUNX1
TGASTCA	7.89%	5.09%	$p < 0.0007$	AP-1
ACAGGAARY	11.57%	8.48%	$p < 0.0013$	ELF1
AGAGGAAGTG	6.41%	4.21%	$p < 0.0017$	PU.1

Extended Data Figure 9 | Motifs directly altered by or adjacent to candidate causal SNPs. a, Known motifs (identified by conservation or SELEX) created or disrupted by candidate causal SNPs at a higher frequency than expected by chance when compared to control SNPs drawn from the same loci.

b, Additional motifs, identified by conservation, created or disrupted by candidate causal SNPs more frequently than by chance. c, Known motifs significantly enriched within 100 bp of candidate causal SNPs, compared to background control SNPs drawn from the same loci.



Extended Data Figure 10 | Enrichment of candidate causal eQTL SNPs in functional elements. **a**, PICS was used to identify candidate causal SNPs for 4,136 eQTL signals in peripheral blood. Bar plot show their overlap with indicated functional genic annotations. Background expectation was calculated based on frequency-matched control SNPs drawn from within 50 kb of the

candidate causal SNPs. **b**, Overlap of candidate causal eQTL SNPs with immune enhancers and promoters, versus background. **c**, Magnitudes of disease-associated eQTLs compared to the space of all eQTLs. Histogram compares the magnitudes of PICS eQTL SNPs that overlap PICS autoimmunity SNPs against the full set of PICS eQTL SNPs.

Transcription factor binding dynamics during human ES cell differentiation

Alexander M. Tsankov^{1,2,3}, Hongcang Gu¹, Veronika Akopian^{2,3}, Michael J. Ziller^{1,2,3}, Julie Donaghey^{1,2,3}, Ido Amit^{1,4}, Andreas Gnirke¹ & Alexander Meissner^{1,2,3}

Pluripotent stem cells provide a powerful system to dissect the underlying molecular dynamics that regulate cell fate changes during mammalian development. Here we report the integrative analysis of genome-wide binding data for 38 transcription factors with extensive epigenome and transcriptional data across the differentiation of human embryonic stem cells to the three germ layers. We describe core regulatory dynamics and show the lineage-specific behaviour of selected factors. In addition to the orchestrated remodelling of the chromatin landscape, we find that the binding of several transcription factors is strongly associated with specific loss of DNA methylation in one germ layer, and in many cases a reciprocal gain in the other layers. Taken together, our work shows context-dependent rewiring of transcription factor binding, downstream signalling effectors, and the epigenome during human embryonic stem cell differentiation.

Human embryonic stem (ES) cells hold great promise for tissue engineering and disease modelling; yet a key challenge to deriving mature, functional cell types is understanding the molecular mechanisms that underlie cellular differentiation.

There has been much progress in understanding how core regulators such as OCT4 (also known as POU5F1), SOX2, and NANOG as well as transcriptional effector proteins of signalling pathways, such as SMAD1, TCF3, and SMAD2/3, control the molecular circuitry that maintains human ES cells in a pluripotent state^{1,2}. While the genomic binding sites of many of these factors have also been mapped in mouse ES cells, cross-species comparison of OCT4 and NANOG targets showed that only 5% of regions are conserved and occupied across species³. Together with more general assessment of divergent transcription factor (TF) binding⁴, those results highlight the importance of obtaining binding data in the respective species.

It is well understood that epigenetic modifications, such as DNA methylation and posttranslational modifications of the various histone tails, are essential for normal development^{5,6}. TF binding sites are overlapping with regions of dynamic changes in DNA methylation and are linked to its targeted regulation^{7,8}. More generally, TFs orchestrate the overall remodelling of the epigenome, including the priming of loci that will change expression only at later stages^{6,9,10}. It has also been shown that lineage-specific TFs and signalling pathways collaborate with the core regulators of pluripotency to exit the ES cell state and activate the transcriptional networks governing cellular specification^{11,12}. However, how the handoff between the central regulators occurs and what role individual TFs and signalling cues play in rewiring the epigenome to control proper lineage specification and stabilize commitment is still poorly understood.

TF binding maps across human ES cell differentiation

To dissect the dynamic rewiring of TF circuits, we used human ES cells to derive early stages of endoderm (dEN), mesoderm (dME) and ectoderm (dEC)^{13–15} along with a mesendoderm (dMS) intermediate (Fig. 1a, Supplementary Information). We defined and collected the dMS population at 12 h owing to maximal expression of *BRACHYURY (T)* (Fig. 1b),



EPIGENOME ROADMAP
A Nature special issue
nature.com/epigenomeroadmap

and carried out chromatin immunoprecipitation followed by sequencing (ChIP-seq) for four of the Roadmap Epigenomics Project¹⁶ core histone modifications (H3K4me1, H3K4me3, H3K27Ac and H3K27me) as well as RNA sequencing (RNA-seq) of polyadenylated transcripts (Supplementary Table 1).

As expected, we observe upregulation of key TFs including *FOXA2* and *HNF4A* in dEN, *HAND1* and *SNAIL2* in dME, and *OTX2* and *PAX6* in dEC (Fig. 1b,c)^{9,17}. We identified high-quality antibodies for 38 factors (Fig. 1c) and provide detailed information, including their validation and use in other studies, in Supplementary Table 2.

Using a micrococcal nuclease (MNase)-based ChIP-seq (MNChIP-seq) protocol¹⁸ we obtained binding patterns as well as reproducibility comparable to sonication ChIP-seq with only 1–2 million cells (Extended Data Fig. 1a–e). We quantified the enrichment over background for each experiment (Supplementary Table 3), and show that the level of binding is comparable to TF ChIP-seq data from ENCODE¹⁹ (Extended Data Fig. 1f). To evaluate computationally the specificity of the chosen antibodies we searched our binding maps for previously reported motifs of the respective factors²⁰ (Extended Data Fig. 2). Our final data set consists of 6.7 billion aligned sequencing reads that yield 4.2 million total binding events (Supplementary Table 3). The binding spectrum of all TFs averages 21,468 peaks and ranges from 578 to 100,778 binding events. Of these 23% are found in promoters, 44% in distal regions, 30% in introns, and 3% in exons.

Classes of TF dynamics

We first grouped the TF binding dynamics into four main classes (static, dynamic, enhanced and suppressed) similar to prior studies in yeast²¹ and then further subdivided each of these as either temporal (between successive time-points) or cross-lineage (between germ layers) (Fig. 2a and Extended Data Figs 3 and 4).

A number of factors, including NANOG, show largely static binding in ES cells and endoderm (Fig. 2a). This could be the result of NANOG's proposed functions in endoderm, including protection against neuroectoderm specification and buffering TGF- β signalling to avoid premature induction of definitive endoderm¹¹. CTCF is both temporally

¹Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA. ²Harvard Stem Cell Institute, Cambridge, Massachusetts 02138, USA. ³Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, Massachusetts 02138, USA. ⁴Department of Immunology, Weizmann Institute, Rehovot, 76100 Israel.

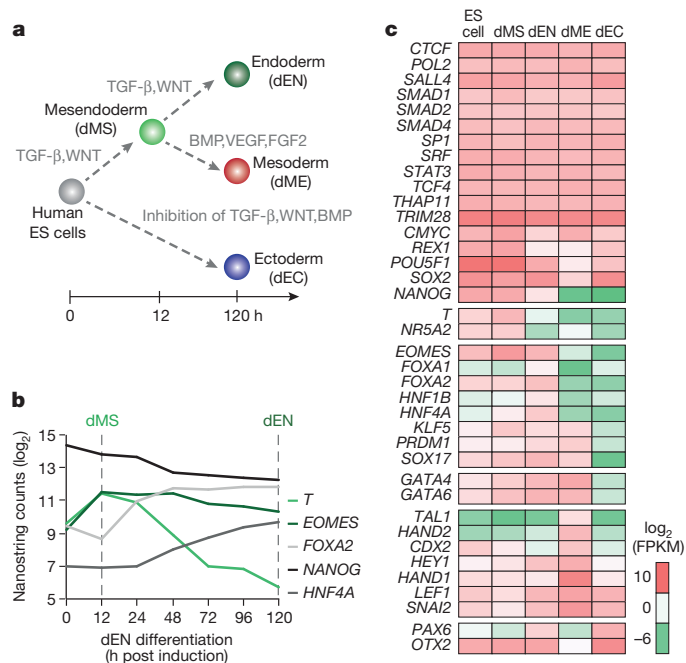


Figure 1 | TF dynamics during human ES cell differentiation. **a**, Schematic of the human ES cell differentiation system including timeline and key signalling pathways that are modulated. **b**, Normalized RNA expression of selected TFs over the differentiation timeline towards endoderm. **c**, RNA-seq data of the selected TFs. Factors are generally ordered by condition where they are most active: ES cells on top, followed by dMS, dEN, dME, and dEC.

and cross-lineage static in its binding pattern, showing a similar overlap between cell types as between replicates (Extended Data Figs 1a and 4a). The high similarity in binding is consistent with a previous study that investigated CTCF binding in 19 diverse human cell types²². Although each of the germ layer derivatives exhibits unique expression signatures, they show overall only limited transcriptional dynamics⁹, which is in agreement with the largely static enrichment for POLII and cMYC (Extended Data Fig. 3a).

In contrast, a number of the selected factors show dynamic binding between two (for example, GATA4) or more (for example, SMAD4) cell types (Fig. 2a, b). EOMES changes its binding profile notably during the dMS to dEN transition, suggesting its function may evolve at different stages of differentiation (Fig. 2c). Also, OTX2 occupies a largely different binding spectrum in the undifferentiated cells compared to dEN and dEC (Fig. 2d). Many factors also exhibit different temporal and cross-lineage dynamics. For example, while NANOG binding is temporally static in dMS and dEN, it is suppressed temporally and cross-lineage in dME (Extended Data Figs 3a, 4b). Meanwhile, OCT4 and SOX2 binding is temporally static in dEN, but cross-lineage dynamic between dEN and dME (Extended Data Figs 3a and 4c). Likewise, TCF4 (a transcriptional effector of WNT signalling) is temporally static in dEN but suppressed in dME and dEC, consistent with the lack of WNT signalling in those germ layers^{13–15} (Extended Data Figs 3a and 4d). Finally, OTX2 is temporally suppressed in dME (Fig. 2a), but temporally dynamic in the other germ layers (Fig. 2d).

To investigate the interplay between TFs across the cell types and how they might collaborate to mediate cellular transitions, we analysed all pairwise TF co-binding relationships. We identify several germ-layer-specific co-binding interactions; for example, GATA4 targets associate significantly (hypergeometric $P < 10^{-300}$) with SMAD1 binding in dME, but less so in dEN (Fig. 3a, left and Extended Data Fig. 5). To extend this, we clustered all co-binding relationships and identified groups of interactions between factors and developmental time points (Fig. 3a, right). We found both clusters of many regulators in one cell type as well as clusters for individual TFs across cell types. For instance, cluster C1 shows

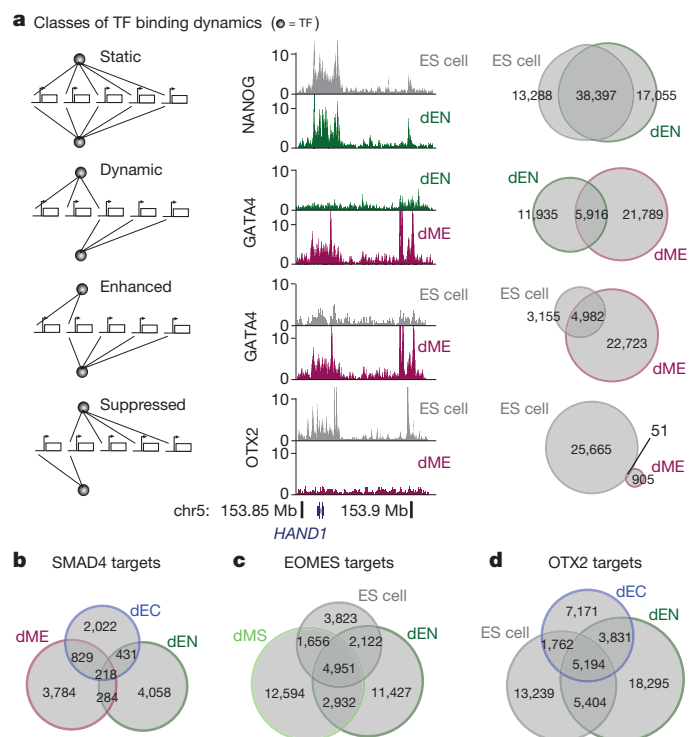


Figure 2 | Classes of TF binding dynamics in germ layers. **a**, Classes of dynamics comparing TF binding between successive time points (temporal) or between different germ layers (cross-lineage). The schematics, browser images, and Venn diagrams illustrate examples of each class. **b**, SMAD4 predominantly binds to unique regions in the three germ layers. **c**, EOMES binding is enhanced from ES cells to dMS and dynamic in dEN. **d**, OTX2 binding is dynamic in dEN and dEC when compared to ES cells.

that CTCF binding spectrum is highly similar in all three germ layers. In cluster C2, we find high overlap in binding between key endoderm regulators, while C4 captures primarily pluripotent and dMS binding profiles. Many known mesoderm factors aggregate in clusters C8 and flanking the pluripotent cluster C4 are EOMES, T, and NR5A2 clusters (C3, C5, C6), all known regulators in mesoderm that are likely to be involved in the transition towards mesoderm and endoderm¹¹.

Interestingly, we noticed that GATA4 and OTX2 binding in the different cell types is not only divergent, but enriched at distinct genomic features (Fig. 3b). In dME 36% of all GATA4 binding sites occur in promoters, compared to only 13.6% in dEN. OTX2's fraction of binding sites at promoters is larger in dEN (34%) and dEC (28%) than in ES cells (13%). Accompanying GATA4's shift in binding preference, we also observe higher levels of H3K4me1 at dEN targets and higher H3K27Ac and H3K4me3 enrichment in dME (Fig. 3c). Similarly, OTX2 associates with higher H3K27Ac and H3K4me1 levels in ES cells, and higher H3K4me3 occupancy in dEN and dEC, in line with increased promoter binding in these two germ layers (Fig. 3c). It is worth noting that similar to the distinct GATA4/SMAD1 co-binding, OTX2 co-occupies a higher fraction of loci with SMAD1 in dEN than in dEC (Fig. 3a, left and Extended Data Fig. 5). Although TGF- β signalling is primarily associated with effector proteins SMAD2/3, it also acts through the SMAD1/5/8 complex and may encourage interaction with OTX2 in dEN but not in dEC, where TGF- β signalling is specifically inhibited²³.

H3K27Ac domains identify lineage regulators

Extended H3K27Ac domains have recently been termed super-enhancers and were used to describe regulatory regions that enrich for binding sites of master TFs in the respective cell types^{24,25}. Binding of GATA4 in dME indeed coincides with long stretches of H3K27Ac near several mesodermal genes (Fig. 4a). We therefore used the previously described

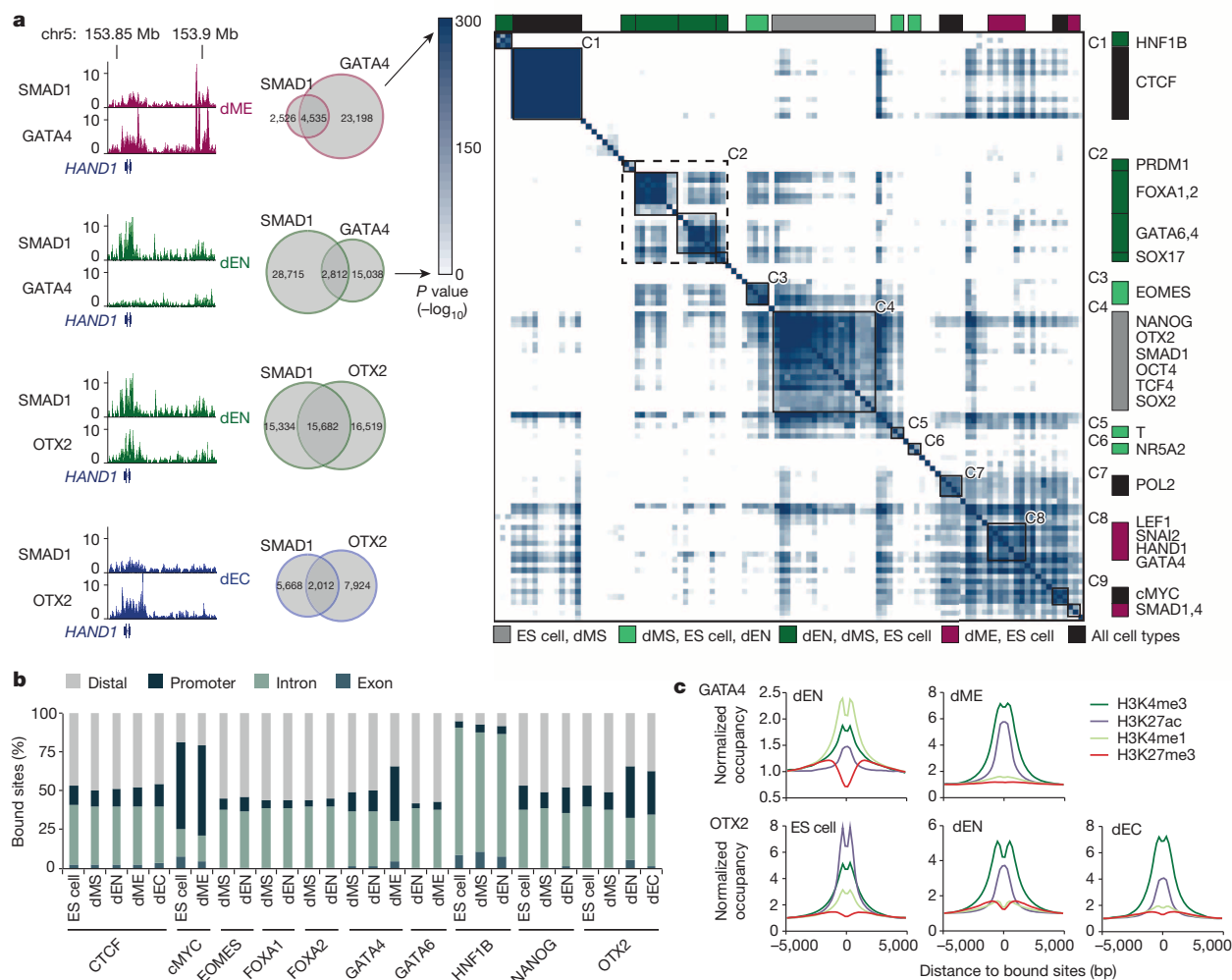


Figure 3 | TF co-binding relationships and genomic targets. **a**, Left, overlap in binding between GATA4 and SMAD1 is greater in dME than in dEN. Similarly, overlap in binding between OTX2 and SMAD1 is greater in dEN than in dEC. Right, highly significant TF co-binding relationships are assigned a dark blue colour, representing $-\log_{10}$ of hypergeometric P value. All TF dynamics and co-binding interactions are clustered and displayed in a matrix,

where each row/column represents a single ChIP-seq experiment. The colour code indicates the cell type identity for the majority of ChIP-seq profiles making up each cluster. **b**, Genomic annotations for factors that bind more than 15,000 regions in multiple conditions. **c**, GATA4 (top) and OTX2 (bottom) binding is associated with different chromatin marks between lineages.

approach^{24,25} to rank extended H3K27Ac domains in our populations and identify such super-enhancers (Supplementary Table 4), which were indeed predominantly unique to each cell type (Fig. 4b and Extended Data Fig. 6). As expected, in human ES cells, core regulators OCT4, SOX2, NANOG (abbreviated OSN), and OTX2 binding is highly enriched at super-enhancers^{1,26} (Fig. 4c).

We used enrichment of binding at super-enhancers for identifying possible master regulators in the germ layers (Fig. 4c); the results were highly robust to different cut-offs for defining the super-enhancers (Supplementary Table 5). Surprisingly, we found that many of the core regulators bound at ES cell super-enhancers also occupy dEN super-enhancers, including OSN, OTX2, SMAD1, TCF4, and SMAD2/3 (Fig. 4c and Extended Data Fig. 6e). In mesoderm, GATA4 and SMAD1 were the most highly enriched factors at dME super-enhancers (Extended Data Figs 6f and 7), consistent with GATA4's known role in directing cardiomyocyte development downstream of BMP signalling²⁷. OTX2 is known to regulate neuronal subtype specification in the midbrain²⁸ and we found strong enrichment for OTX2 binding at ectoderm super-enhancers (Fig. 4c and Extended Data Fig. 6g, h). Meanwhile, dMS super-enhancers were enriched for known regulators such as EOMES and T, along with OSN and OTX2 (Fig. 4c). At a lower significance level we also find enrichment for a number of endoderm factors, including FOXA1/2,

GATA4/6 and SOX17 (Supplementary Table 5). Interestingly, binding of EOMES, T and FOXA1/2 in the undifferentiated ES cells was also enriched (hypergeometric $P < 10^{-6}$) at dMS super-enhancers (Fig. 4c and Extended Data Fig. 6), suggesting that a number of loci might be already marked before differentiation.

Regulation of poised enhancers across germ layers

As dEN H3K27Ac domains were mostly devoid of known endoderm TFs, we asked if such regulators are instead present at regions that enrich for H3K4me1, as seen at the *HNF1B* locus (Fig. 5a). H3K4me1 can be found at both active and poised enhancers²⁹ and is known to also form extended enhancer domains that may not overlap with the H3K27Ac domains^{24,25}. Using the same approach as above we identified extended H3K4me1 domains in dEN and then measured enrichment for TF binding in these regions. In contrast to H3K27Ac, the top H3K4me1 domains were enriched for binding of FOXA1/2, GATA4, GATA6, and SOX17 (Extended Data Fig. 8a, b), known regulators of the early endodermal fate³⁰. We then measured the significance in overlap between TF binding and all poised enhancers for each cell type and found strong enrichment for these regulators and PRDM1 in dEN (Extended Data Fig. 8c, d).

In concordance with this analysis and global chromatin remodelling trends (Extended Data Fig. 8e), GATA4 is associated with dynamics of

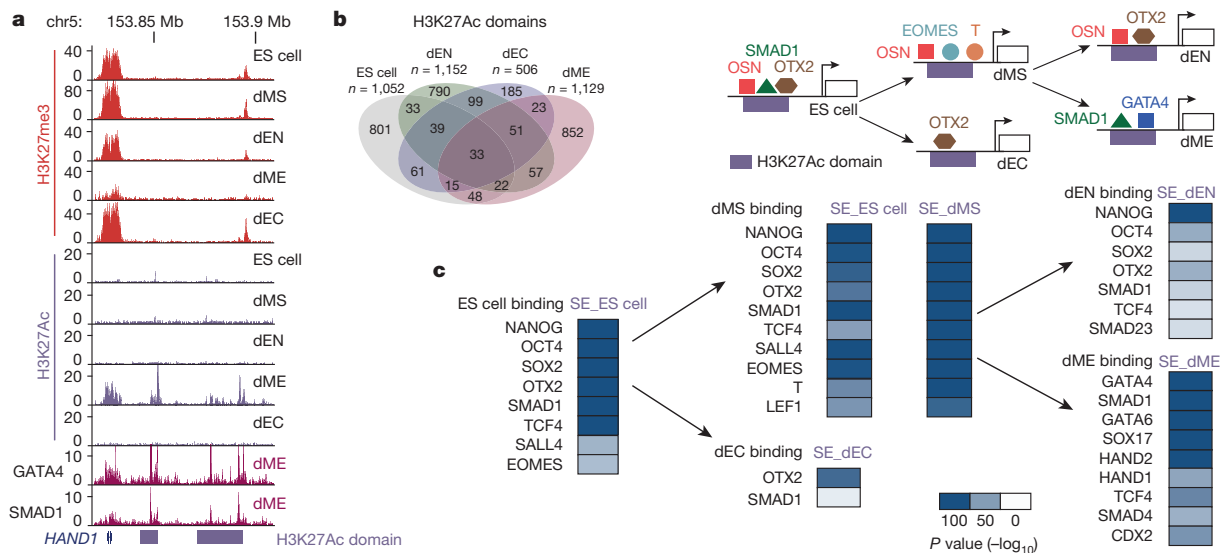


Figure 4 | Extended H3K27Ac domains highlight unique TF transitions. **a**, Browser tracks for H3K27me3 and H3K27Ac across all five cell types as well as GATA4/SMAD1 enrichment over the *HAND1* locus in dME. **b**, Limited overlap of extended H3K27Ac domains between cell types. **c**, Top, schematic of

different transitions in TF regulation at super-enhancers. OTX2 is present at the dMS stage, but not shown in the schematic due to limited space. Bottom, hypergeometric P values ($-\log_{10}$) displaying the most significant overlaps in H3K27Ac super-enhancers (SE) and TF binding for each cell type.

H3K4me1 in dEN and H3K27Ac in dME. Given that the SMAD proteins are known to interact with histone acetyltransferases EP300 and CBP³¹, it is plausible that, through BMP signalling in dME, GATA4 interacts with SMAD1 and recruits EP300 to induce acetylation of H3K27 at target sites. This recruitment relationship is further supported by the higher enrichment of GATA4 motif instances at SMAD1 binding sites in dME versus dEN (Fig. 5b and Extended Data Fig. 8f) and the stronger enrichment of H3K27Ac at GATA4 targets in dME versus dEN (Fig. 3c).

To further explore this, we used several shRNAs to knock down (KD) GATA4 and then measured gene expression following differentiation into dME and dEN (Extended Data Fig. 9a). The mean expression for more than 20 lineage markers is very similar between control and KD cell lines, arguing that the KD cells still differentiate into comparable

populations (Fig. 5c, right bar). While the GATA4 KD in dEN does not greatly affect any of the measured endoderm TFs (total $P = 0.49$, paired t -test), in dME the KD leads to a 1.7–4-fold reduction in the expression of seven key factors (total $P = 5.39 \times 10^{-5}$, paired t -test). GATA4 binding in dME and dEN occupies similar loci in control and KD cell lines (Extended Data Fig. 9b, c), and H3K27Ac super-enhancers in dME are largely unaffected by our knockdown (Extended Data Fig. 9d, e). Nonetheless, we observe a significant decrease in SMAD1 and H3K27Ac enrichment in dME at GATA4 target sites in the KD lines (Fig. 5d, $P < 10^{-300}$, paired t -test). To a lesser degree, we also observe a decrease in mean SMAD1 occupancy at binding sites away from GATA4 (Extended Data Fig. 9f). This could be the result of the general reduction of SMAD1 binding, such as factors from the TEAD and GATA family (Fig. 5b).

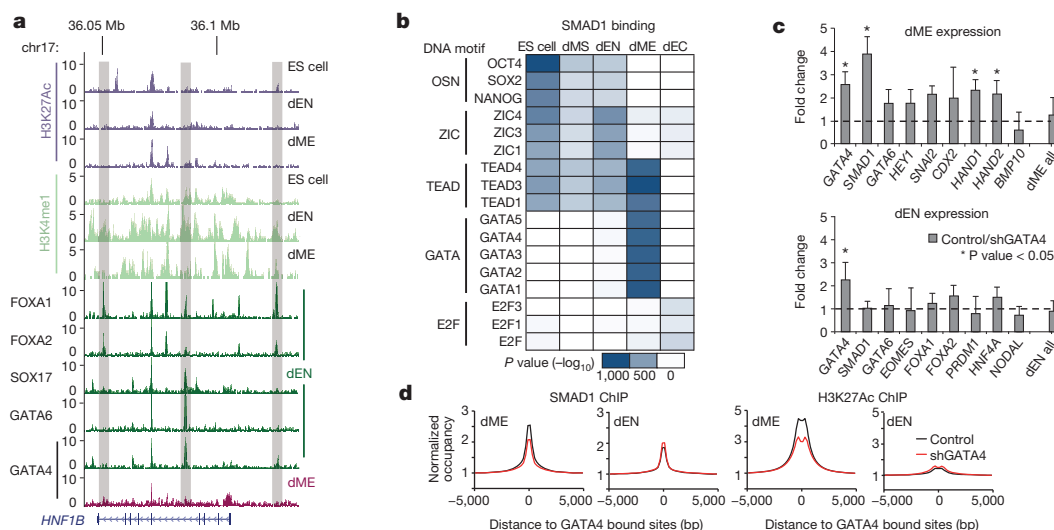


Figure 5 | Regulatory dynamics at putative poised enhancers. **a**, Selected browser tracks for H3K27Ac and H3K4me1 and normalized binding of FOXA1/2, SOX17, and GATA4/6 over the *HNF1B* locus. Grey vertical bars highlight regions enriched for H3K4me1 in dEN. **b**, P values ($-\log_{10}$) for three or more of the most enriched DNA binding motifs (rows) at SMAD1 binding per cell type (columns). **c**, Gene expression of selected lineage markers in dME (top) and dEN (bottom), comparing three GATA4 KD and control lines,

determined by reverse transcription quantitative PCR (RT-qPCR). The mean expression for 22 dEN and 24 dME marker genes (excluding *GATA4* and *SMAD1*) is shown as the last bar in each panel. Error bars display the standard deviation in fold expression change. Asterisk highlights genes with significant ($P < 0.05$, paired t -test) change in expression between control and KD replicates. **d**, Normalized SMAD1 (left) and H3K27Ac (right) occupancy decreases in shRNA KD versus control lines in dME but not in dEN.

Loss of DNA methylation at targets of lineage TFs

DNA methylation can silence genomic regions, directly or indirectly, and plays an important role during mammalian development⁵. Some TFs can modulate DNA methylation levels⁸, but it is not generally known which factors can alter it in a developmental context and which ones might be sensitive to its presence. In endoderm at a region upstream of *SOX17*, we observe specific loss of DNA methylation accompanied by epigenetic remodelling to a poised state. We also observe that the loss of DNA methylation associates with lineage-specific binding of several TFs (Fig. 6a and Extended Data Fig. 10a). Interestingly, OTX2 and NANOG show some enrichment already in ES cells that seems to be linked to a very focal depletion of DNA methylation that may serve as a means of initial marking or protecting the region for downstream binding (Extended Data Fig. 10b).

We next performed global enrichment analysis for all TF binding at regions that either gained or lost DNA methylation. Many target sites of OSN as well as SMAD1 and TCF4 show gain of DNA methylation in all three lineages, consistent with silencing of their pluripotency-related

target genes (Fig. 6b, left). The dMS target sites of T and EOMES also become methylated in the three germ layer populations. Interestingly, we frequently find a reciprocal gain in DNA methylation in the alternative lineages of key dEN and dEC factors (Fig. 6b, middle).

As shown near *SOX17*, we also find that lineage regulators associate with targeted loss of DNA methylation. For instance, in dEN binding sites of *EOMES*, *FOXA1/2* (Extended Data Fig. 10c, d), *GATA4/6*, *SOX17*, and *OTX2* display focal and germ layer specific loss of DNA methylation (Fig. 6b, c). We also find strong enrichment for loss of DNA methylation at *OTX2* binding sites in dEC (Fig. 6b, d). In dME we find seven partially overlapping TFs that show loss of DNA methylation at their binding sites, especially in regions that also gain H3K27Ac (Fig. 6b, e and Extended Data Fig. 7c). Using reduced-representation bisulfite sequencing³² we measured the DNA methylation level for a representative subset of targets in *GATA4* KD and control lines. Both dME and dEN *GATA4* KD cells displayed significantly higher methylation level ($P < 10^{-10}$, paired *t*-test) (Fig. 6f and Extended Data Fig. 10e), suggesting a possible role for *GATA4* in the focal depletion of DNA methylation.

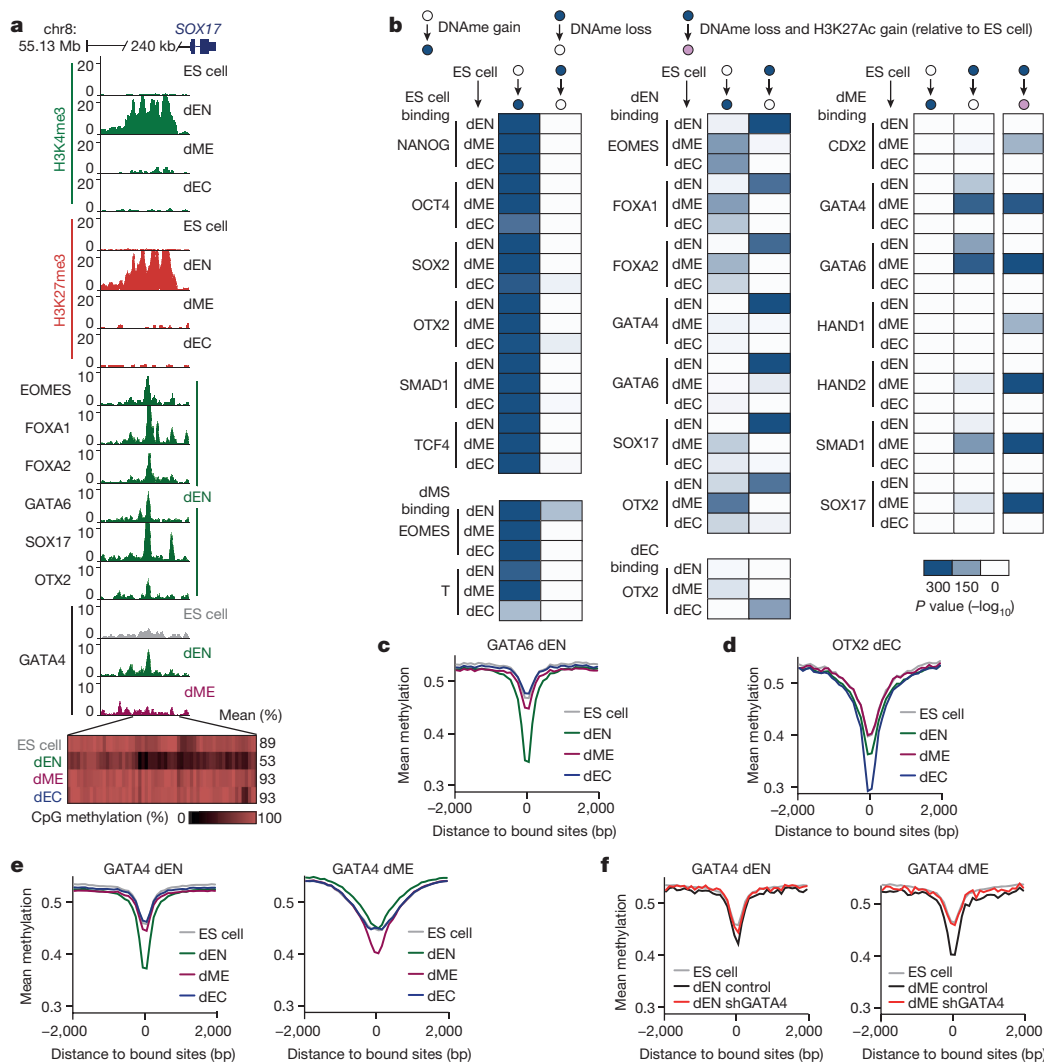


Figure 6 | Specific loss of DNA methylation at targets of key lineage TFs.

a, Top, browser tracks for H3K4me3 and H3K27me3 as well as enrichment of selected TFs upstream of *SOX17*. Bottom, each rectangle represents a single CpG and its methylation state. Loss of DNA methylation occurs specifically in dEN, which coincides with changes in chromatin state and specific binding of several known endoderm factors. **b**, Hypergeometric *P* values ($-\log_{10}$) for the overlap in TF binding and regions that gain or lose DNA methylation (DName) relative to ES cells. Possible transition states are defined at the top. Heat maps display the enrichment of TF binding in ES cells, dMS (left), dEN,

dEC (centre), and dME (right) at differentially methylated regions in the three germ layers. **c**, Whole-genome bisulfite sequencing (WGBS)-based average CpG methylation level of 100-bp tiles over GATA6-bound dEN targets.

d, WGBS mean methylation level at OTX2 dEC targets. **e**, WGBS mean methylation level at GATA4 dEN and dME targets. **f**, Reduced-representation bisulfite sequencing-based average CpG methylation level of 100-bp tiles over GATA4 targets in control and GATA4 KD cell lines in dEN (left) and dME (right). For comparison, WGBS ES cell mean methylation level is also shown (grey).

Discussion

Directed differentiation of human ES cells into the three embryonic germ layers coupled with comprehensive TF binding analysis and integration with epigenomic data has allowed us to characterize differentiation-associated regulatory dynamics. We find that targets of many lineage-specific factors associate with loss of DNA methylation in those germ layers, while factors that are expressed in more than one lineage (GATA4, GATA6, OTX2, SOX17) show a corresponding loss of DNA methylation at their targets in multiple cell types. This is in line with the model that some TFs have an intrinsic ability to alter DNA methylation, although more work is needed to determine if all of these can indeed be considered “pioneer factors”³³. We also find a specific gain of DNA methylation for the targets of many TFs at later time points or in parallel time-points but along alternate lineages. This might present a possible mechanism for occluding binding sites of certain methylation sensitive factors at past or alternate differentiation paths.

To investigate the interplay between TF binding and the chromatin landscape, we focused on TF dynamics at H3K27Ac super-enhancers, where OTX2 and OSN seem to guide the transition to dEN while GATA4 and OTX2 act as key regulators for dME and dEC, respectively. GATA4 exemplifies a factor with distinct germ layer functions, where in dEN it resides at poised enhancers and in dME it appears to associate with SMAD1/EP300 to establish and maintain H3K27Ac domains. The dual use of GATA4 and OTX2 highlights the modularity in transcriptional networks in development and the complex interaction of downstream signalling effectors, TFs and chromatin in the three germ layers.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 5 December 2013; accepted 14 January 2015.

- Boyer, L. A. *et al.* Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122**, 947–956 (2005).
- Young, R. A. Control of the embryonic stem cell state. *Cell* **144**, 940–954 (2011).
- Kunars, G. *et al.* Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nature Genet.* **42**, 631–634 (2010).
- Villar, D., Flicek, P. & Odom, D. T. Evolution of transcription factor binding in metazoans - mechanisms and functional implications. *Nature Rev. Genet.* **15**, 221–233 (2014).
- Smith, Z. D. & Meissner, A. DNA methylation: roles in mammalian development. *Nature Rev. Genet.* **14**, 204–220 (2013).
- Cantone, I. & Fisher, A. G. Epigenetic programming and reprogramming during development. *Nature Struct. Mol. Biol.* **20**, 282–289 (2013).
- Ziller, M. J. *et al.* Charting a dynamic DNA methylation landscape of the human genome. *Nature* **500**, 477–481 (2013).
- Stadler, M. B. *et al.* DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* **480**, 490–495 (2011).
- Gifford, C. A. *et al.* Transcriptional and epigenetic dynamics during specification of human embryonic stem cells. *Cell* **153**, 1149–1163 (2013).
- Lara-Astiaso, D. *et al.* Immunogenetics. Chromatin state dynamics during blood formation. *Science* **345**, 943–949 (2014).
- Teo, A. K. *et al.* Pluripotency factors regulate definitive endoderm specification through eomesodermin. *Genes Dev.* **25**, 238–250 (2011).
- Thomson, M. *et al.* Pluripotency factors in embryonic stem cells regulate differentiation into germ layers. *Cell* **145**, 875–889 (2011).
- Lee, G., Chambers, S. M., Tomishima, M. J. & Studer, L. Derivation of neural crest cells from human pluripotent stem cells. *Nature Protocols* **5**, 688–701 (2010).
- Hay, D. C. *et al.* Highly efficient differentiation of hESCs to functional hepatic endoderm requires ActivinA and Wnt3a signaling. *Proc. Natl Acad. Sci. USA* **105**, 12301–12306 (2008).
- Evseenko, D. *et al.* Mapping the first stages of mesoderm commitment during differentiation of human embryonic stem cells. *Proc. Natl Acad. Sci. USA* **107**, 13742–13747 (2010).
- Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* <http://dx.doi.org/10.1038/nature14248> (this issue).
- Xie, W. *et al.* Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell* **153**, 1134–1148 (2013).
- Henikoff, J. G., Belsky, J. A., Krassovsky, K., MacAlpine, D. M. & Henikoff, S. Epigenome characterization at single base-pair resolution. *Proc. Natl Acad. Sci. USA* **108**, 18318–18323 (2011).
- Gerstein, M. B. *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91–100 (2012).
- Jolma, A. *et al.* DNA-binding specificities of human transcription factors. *Cell* **152**, 327–339 (2013).
- Harbison, C. T. *et al.* Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**, 99–104 (2004).
- Wang, H. *et al.* Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Res.* **22**, 1680–1688 (2012).
- Chambers, S. M. *et al.* Highly efficient neural conversion of human ES and iPS cells by dual inhibition of SMAD signaling. *Nature Biotechnol.* **27**, 275–280 (2009).
- Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934–947 (2013).
- Whyte, W. A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307–319 (2013).
- Buecker, C. *et al.* Reorganization of enhancer patterns in transition from naive to primed pluripotency. *Cell Stem Cell* **14**, 838–853 (2014).
- Pikkariainen, S., Tokola, H., Kerkela, R. & Ruskoaho, H. GATA transcription factors in the developing and adult heart. *Cardiovasc. Res.* **63**, 196–207 (2004).
- Vernay, B. *et al.* Otx2 regulates subtype specification and neurogenesis in the midbrain. *J. Neurosci.* **25**, 4856–4867 (2005).
- Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279–283 (2011).
- Zaret, K. S. Genetic programming of liver and pancreas progenitors: lessons for stem-cell differentiation. *Nature Rev. Genet.* **9**, 329–340 (2008).
- Pouponnot, C., Jayaraman, L. & Massagué, J. Physical and functional interaction of SMADs and p300/CBP. *J. Biol. Chem.* **273**, 22865–22868 (1998).
- Meissner, A. *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454**, 766–770 (2008).
- Zaret, K. S. & Carroll, J. S. Pioneer transcription factors: establishing competence for gene expression. *Genes Dev.* **25**, 2227–2241 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements We would like to thank all members of the Meissner laboratory for their support and feedback. We also thank F. Kelley and other members of the Broad Technology Labs and Sequencing Platform as well as J. Doench and members of the Genetic Perturbation Platform at the Broad Institute. We would like to thank L. Gaffney for graphical support. This work was supported by the NIH Common Fund (U01ES017155), NIGMS (P01GM099117), NHGRI (P50HG006193) and the New York Stem Cell Foundation. A.M.T. was supported by NIH Ruth L. Kirschstein NRSA fellowship 5F32DK095537. A.M. is a New York Stem Cell Foundation Robertson Investigator.

Author Contributions A.M.T. and A.M. designed and conceived the study. A.M.T. performed the experiments and all analysis, H.G. generated libraries with supervision from A.G., V.A. performed cell culture, M.J.Z. helped with data processing and analysis, J.D. performed experiments, I.A. provided experimental advice, A.M.T. and A.M. interpreted the data and wrote the manuscript.

Author Information All data have been deposited in GEO under accession code GSE61475. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.M. (alexander_meissner@harvard.edu).



This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported licence. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons licence, users will need to obtain permission from the licence holder to reproduce the material. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-sa/3.0>

METHODS

No statistical methods were used to predetermine sample size.

Human ES cell culture. Cell culture was done as reported previously⁹. Briefly, we chose the NIH approved, male human embryonic stem (ES) cell line HUES64 because it has maintained a stable karyotype over many passages and is able to differentiate well into the three germ layers. HUES64 was routinely tested for Mycoplasma and was negative in all instances. ES cells were maintained on $\sim 15,000$ cells per cm^2 irradiated murine embryonic fibroblasts (MEFs, Global Stem) and cultured in 20% knockout serum replacement (KSR, Life Technologies), 200 mM GlutaMAX (Life Technologies), $1\times$ minimal essential media (MEM) non-essential amino acids solution (Life Technologies), $10\text{ }\mu\text{g ml}^{-1}$ bFGF (Millipore), $55\text{ }\mu\text{M}$ β -mercaptoethanol in knockout Dulbecco's modified Eagle medium (KO DMEM, Life Technologies). ES cells were passaged every 4–5 days using 1 mg ml^{-1} collagenase IV (Life Technologies).

Directed differentiation of human ES cells. When human ES cells reached 60–70% confluency on MEFs, the cells were plated as clumps on 6-well plates coated with Matrigel (Life Technologies) in mTeSR1 basal medium (Stem Cell Technologies). We maintained the cells for three days in feeder-free culture and then induced directed differentiation towards mesendoderm, endoderm, mesoderm, and ectoderm using different media conditions. For mesendoderm and endoderm differentiation cells were cultured for 12 and 120 h, respectively, in Roswell Park Memorial Institute (RPMI) medium (Life Technologies) supplemented with 100 ng ml^{-1} activin A (R&D Systems), 50 nM ml^{-1} WNT3A (R&D Systems), 0.5% FBS (Hyclone), 200 mM GlutaMAX (Life Technologies), $0.2\times$ MEM non-essential amino acids solution (Life Technologies), and $55\text{ }\mu\text{M}$ β -mercaptoethanol. For the first 24 h of mesoderm differentiation, cells were cultured in DMEM/F12 medium supplemented with 100 ng ml^{-1} activin A (R&D Systems), 10 ng ml^{-1} bFGF (Millipore), 100 ng ml^{-1} BMP4 (R&D Systems), 100 ng ml^{-1} VEGF (R&D Systems), 0.5% FBS (Hyclone), 200 mM GlutaMAX (Life Technologies), $0.2\times$ MEM non-essential amino acids solution (Life Technologies), and $55\text{ }\mu\text{M}$ β -mercaptoethanol. From 24 to 120 h of mesoderm differentiation, Activin A was removed from the culture. For ectoderm differentiation cells were cultured in DMEM/F12 medium supplemented with $2\text{ }\mu\text{M}$ TGF- β inhibitor (Tocris, A83-01), $2\text{ }\mu\text{M}$ WNT3A inhibitor (Tocris, PNU-74654), $2\text{ }\mu\text{M}$ dorsomorphin BMP inhibitor (Tocris), 15% KOSR (Life Technologies), $0.2\times$ MEM non-essential amino acids solution (Life Technologies), and $55\text{ }\mu\text{M}$ β -mercaptoethanol. Media was changed daily. Before inducing differentiation, we manually removed the differentiated cell clumps. We routinely obtain greater than 80% differentiated cells based on the presence of the surface marker CD56 (81.7% of mesoderm and 94.4% of ectoderm cells) and greater than 70% differentiated cells based on the surface marker CD184 for endoderm.

RNA extraction and RNA-seq. For measuring expression levels, RNA was isolated from the human ES cells and differentiated cells using TRIzol (Invitrogen, 15596-026), further purified with RNeasy columns (QIAGEN, 74104) and DNase treated. RNA-seq library construction and data analysis was carried out as described previously⁹.

Antibodies. Supplementary Table 2 lists detailed information for all antibodies used in this study, along with references that validate the specificity and use of each antibody.

MNChIP-seq and library construction. ChIP-seq for all chromatin marks was done as in ref. 9. MNChIP-seq for TFs was carried out as in ref. 9 with several modifications, including the micrococcal nuclease (MNase) digestion. Briefly, cell were grown to a final count of 10 million, resuspended in PBS, and crosslinked in 10% formaldehyde solution for 10 min at room temperature. Following quenching with 0.125M glycine and two PBS washes, we isolated nuclei using cell lysis buffer (20 mM Tris-HCl pH 8, 85 mM KCl, 0.5% NP40). Nuclei were then digested using MNase (Worthington, LS004797) as done in ref. 18. Digestion was stopped with 0.05 M EGTA and chromatin was aliquoted into 1–2 million cells per ChIP. Antibodies were added and immunoprecipitation was carried out overnight at 4°C as done in ref. 9. The next day, protein G beads (Life Technology, 10009D) were added for 2 h at 4°C to isolate the protein-bound DNA and washed twice using low salt wash buffer (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris-HCl pH 8.1, 150 mM NaCl), high salt wash buffer (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris-HCl pH 8.1, 500 mM NaCl), LiCl wash buffer (0.25 M LiCl, 0.5% NP40, 0.5% sodium deoxycholate, 1 mM EDTA, 10 mM Tris-HCl pH 8.1), and TE buffer pH 8 (10 mM Tris-HCl, pH 8, 1 mM EDTA pH 8). DNA was eluted twice using 100 μl of ChIP elution buffer (1% SDS, 0.1 M NaHCO_3) at 65°C for 15 min. Crosslinking was reversed by addition of 32 μl reverse crosslinking salt mixture (250 mM Tris-HCl pH 6.5, 62.5 mM EDTA pH 8, 1.25 M NaCl, 5 mg ml^{-1} proteinase K) for 5–18 h at 65°C . DNA was isolated using phenol/chloroform extraction and treated with DNase-free RNase for 30 min at 37°C . The whole-cell extract (WCE) control was generated using MNase-treated material that was then reverse-crosslinked and phenol/chloroform-extracted, skipping the immunoprecipitation and washing steps. DNA libraries were constructed using standard Illumina

protocols for blunt-ending, polyA extension, and ligation, except each clean-up step was replaced with phenol/chloroform extractions to preserve small fragments as done in ref. 18. Ligated DNA was then PCR-amplified and gel-size-selected for fragments between 30 and 600 bp. Samples were sequenced using Illumina HiSeq at a target sequencing depth of 20 million uniquely aligned reads.

shRNA infection and knockdown experiments. ES cells were maintained in KSR culture media as described above and passaged onto geltrex coated dishes in mTeSR1 culture media before infection. When cells were $\sim 75\%$ confluent, cells were collected with accutase as single cells or small clumps. 100,000 ES cells were plated per well of 12-well plate coated with geltrex and in mTeSR1 culture media. After 24 h, ES cells were infected twice on separate days for 3 h with approximately 30 viral particles per cell. 48 h after the last infection, cells were selected with $1\text{ }\mu\text{g ml}^{-1}$ puromycin until the non-infected ES cells die off (usually within 3 days). Knockdown (KD) and control shRNA-infected ES cell lines were then maintained as described above. We then performed directed differentiation of three control and KD cell lines into 5-day dEN and dME. We collected cells and carried out RNA and DNA extraction as ref. 9. cDNA reaction was set-up from 1 μg of total RNA per sample using High-Capacity cDNA RT Kit (Life Technologies). qPCR was performed on 384-well TaqMan hPSC Scorecard plates using ViiA7 RUO software and Applied Biosystems ViiA7 instrument. C_T values were normalized using two probes of the ACTN housekeeping gene and averaged for the three GATA4 KD and three control cell lines to obtain fold change in expression. DNA was used for reduced-representation bisulfite sequencing as in ref. 32. We also collected crosslinked cells from the same samples and carried out MNChIP-seq for GATA4, SMAD1, and H3K27Ac as described above. Composite plots display the average normalized occupancy for three GATA4 KD and two control cell lines. We used pLKO.1 cloning vector with the following target sequences for GATA4 KD: CCAGAGATTCTG CAACACGAA, CGAGGAGATGCGTCCCATCAA, CCCGGCTTACATGGCC GACGT. The shRNA control cell lines targeted gene products not present in the human genome using the same cloning vector with the following target sequences: TGACCCTGAAGTTCATCTGCA (GFP) and CACTCGGATATTTGATATGTG (Luciferase).

Selection of transcription factors. Approximately half of the transcription factors (TFs) were chosen because they are known to play an important role in regulation of pluripotent cells or in the transition to mesendoderm (for example, BRACHYURY), endoderm (for example, SOX17), mesoderm (for example, GATA4), and ectoderm (for example, PAX6). Others were chosen computationally based on Nanostring expression analysis and RNA-seq data. Previous work¹² identified that OCT4 and SOX2 play distinct roles in the transition from ES cells to mesendoderm and ectoderm based on differential expression of these TFs in the two lineages. We used a similar approach to computationally identify factors that are differentially expressed in mesoderm and endoderm. Another study showed that temporal upregulation of TFs can be indicative of their importance at specific stages of blood differentiation³⁴. We used this approach to identify factors that were upregulated upon transition to mesendoderm, mesoderm and endoderm and included those as well in the study (see Supplementary Table 2 for additional details on the factors).

ChIP-seq and MNChIP-seq data processing. Reads were aligned to the hg19 reference assembly using bwa version 0.5.7 (ref. 35) with default parameter settings. Subsequently, reads were filtered for duplicates and extended by 200 bp. For visualization, extended reads were summed at each base and normalized for sequencing depth by scaling the y axis to represent cumulative reads per 1 million reads sequenced. This normalization was used for browser and heat map visualizations of the data in all figures. We used MACS³⁶ peak calling algorithm with default settings to identify significant binding events for each TF, excluding duplicate reads. Peaks were additionally discarded if they overlapped with regions that MACS detected as peaks in four different WCE samples. Such regions have been shown to cause false-positive peaks in ChIP-seq data due to unannotated high copy number regions³⁷. Peaks were then annotated according to their proximity to transcription start sites (TSSs) using Homer³⁸. Peaks within exons and introns were annotated first. Then, peaks overlapping a region from $-2,000\text{ bp}$ to $+500\text{ bp}$ of their nearest TSS were annotated as at promoters. Peaks outside of promoters but not in exons or introns were annotated as distal.

Data quality assessment and motif analysis. To quantify enrichment over background in ChIP-seq experiments, we measured the percentage of reads in peaks by counting all unique tags within 1,000 bp regions centred on all binding events, using bedtools multicov function with default parameters. To compare to ENCODE, we downloaded all ($n = 1,410$) TF ChIP-seq profiles with matching peak and raw data (.bam) files from <http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/>, and computed the percentage reads in peaks in the same manner. Since ENCODE data was collected in cell types where the factors are known to be active, for Extended Data Fig. If we excluded all our TF binding profiles for time points where the factors are not highly expressed and expected to be inactive (middle box plot).

To quantify the specificity of our antibodies computationally, we carried out motif analysis that measured the enrichment of 1,887 known DNA binding sequences at 500 bp regions centred on the peaks of each TF using Centrimo³⁹ and Homer³⁸ (Extended Data Fig. 2). For six factors, (POL2, SALL4, T, NR5A2, THAP11, TRIM28) we did not find a reliable DNA-binding motif in the database of 1,887 motifs combining TRANSFAC and Jolma *et al.* data sets²⁰. For the remaining 32 TFs, we found that 88% (28/32) of factors significantly ($P < 10^{-75}$) associate with the known DNA binding motif. Moreover, we carried out *de novo* motif discovery for these factors (using MEME⁴⁰ and Homer³⁸) and show that these motifs are highly similar to the known motifs, further supporting the specificity of these antibodies (Extended Data Fig. 2). For the other 4 factors (SRF, REX1, STAT3, TAL1) of the 32, we believe that either the known motifs in the database do not match the *in vivo* binding affinities for these factors in our cell types or that cross-reactivity of the antibody with other proteins is occurring. To be conservative, we have excluded all these factors from further analyses, figures, and the main manuscript.

The GATA4 and SMAD1 motif enrichment in Extended Data Fig. 8f was also carried out using Centrimo³⁹ with weighted moving average of 50-bp window. Finally, motif enrichment for Fig. 5b was carried out by scanning 1,887 motifs (see above) within 500 bp of binding using Centrimo³⁹ and displaying three or more of the most enriched DNA motifs per cell type.

TF dynamics and co-binding relationships. For quantifying TF dynamics between cell types and co-binding relationships between TFs, peak regions were merged if two peak centres were a distance of 1,000 bp or less, and significance P values were calculated using the hypergeometric distribution and were subsequently corrected for multiple hypothesis testing. For each TF MNChIP in each condition, we calculated a vector of the $-\log_{10} P$ values for interactions with all other experiments. We then clustered all vectors along both rows and columns based on correlation distance using hierarchical clustering algorithm and average linkage (Fig. 3a). We filtered all experiments with no interactions at significance level P value $< 10^{-5}$ for ease of visualization. To define classes of TF binding dynamics, binding was termed enhanced/suppressed if we observed at least a twofold increase/decrease in binding sites between two different conditions. If the binding sites had not decreased/increased twofold between two conditions, we defined the co-binding relationship as static if P value $< 10^{-300}$, and dynamic if P value $> 10^{-300}$.

Defining chromatin state. For differential signal enrichment analysis, we first computed the number of uniquely aligned sequencing tag midpoints for all 1-kb tiles of the genomic black list filtered human genome. Genomic region black lists were obtained from <http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeMapability/wgEncodeDacMapabilityConsensusExcludable.bed.gz>.

For each histone mark and each condition, we then determined all 1-kb tiles significantly enriched over the whole-cell extract (WCE). To that end, we fitted local Poisson models to the read count normalized WCE tag distribution for each 1-kb tile of the human genome⁴¹. Only regions enriched threefold or higher compared to the whole cell extract and significant after correcting multiple testing using the Benjamini–Hochberg method at a q value ≤ 0.05 were retained. In order to identify differentially enriched regions between ES cells and each of the ES-cell-derived populations, we took advantage of a recently published analysis strategy based on mixture models that allows to incorporate replicate information and to correct for differences in IP efficiency and signal to noise ratio⁴². We used the R implementation in the software package *enrich* to first fit a latent Poisson mixture model with two components to each ChIP-seq experiment in order to obtain an estimate of the fraction of reads in the signal component. Next, we used the initial parameter estimates from the latter model to fit a joint Poisson mixture model for each group of biological replicates. Finally, we used the obtained models for each sample group to conduct pairwise comparisons accounting for sequencing depth and differences in IP efficiency. To that end, we made the assumption that the true number of enriched regions between two compared conditions for a given mark or factor is similar and set the p parameter in the *enrich* mix function to 1. Finally, we obtained a list of candidates of differentially enriched regions at an FDR = 0.05 and retained only those regions that exhibited an absolute \log_2 difference ≥ 1.5 in the estimated tile enrichment levels and that were significantly enriched above background according to the first analysis step. Next, we specifically decided to exclude more gradual changes in histone modifications and restricted the set of differentially enriched regions to those that were above background in one but not the other condition in each of the pairwise comparisons: ES cell vs dMS, ES cell vs dEN, ES cell vs dME and ES cell vs dEC. Based on these differential analysis results, we then binarized our ChIP-seq histone modification enrichment matrix. Next, we used this binarized matrix to assign each tile one of 10 states, now also incorporating DNA methylation data. The states were defined as follows (see below for details) with their order recapitulating their precedence: H3K4me3&H3K27me3, H3K4me3, H3K27me3 & H3K4me1, H3K27ac, H3K4me1, H3K27me3, unmethylated region (UMR, where $0\% \leq \text{UMR} \leq 10\%$ methylation), intermediate methylated region (IMR, where $10\% < \text{IMR} \leq 60\%$ methylation), highly methylated region (HMR, where

$60\% < \text{HMR} \leq 100\%$ methylation), none (no detectable histone modification enrichment or DNA methylation data for a given 1-kb tile).

Super-enhancer analysis. Using chromatin data, we defined super-enhancers as in refs 24, 25. Briefly, we used MACS³⁶ peak calling algorithm (default settings, except $-p$ parameter was set to $1e-9$) to detect enrichments in H3K27Ac ChIP-seq data for each cell type. Peaks were then merged if they were within a distance of 12.5 kb. We then ranked the stitched H3K27Ac enriched regions based on the normalized, background-subtracted average read density (in units of reads-per-million-mapped per bp of stitched region). The cutoff for classifying super-enhancers was defined as refs 24, 25, or the point where a line with a slope 1 is tangent to the curve of normalized region signal versus region ranking. The same procedure was used to define H3K4me1 super-enhancers per cell type.

We also used this procedure to find super-enhancers within a more inclusive set of parameters (MACS parameter $-p$ set to $1e-5$ instead of $1e-9$ and stitching distance set to 5 kb instead of 12.5 kb), but found no differences in our conclusions (Supplementary Table 5). We also found no difference when using other cut-offs for defining super-enhancers (top 250, top 500, top 1,000, and top 2,000 enhancer regions, Supplementary Table 5), and found that using a fixed threshold had the advantage of uniformity between cell types in the enrichment analysis. Finally, excluding all enriched regions within 2,500 kb of TSSs also led to highly similar results and did not change our conclusions.

Chromatin states versus super-enhancers. H3K27Ac chromatin states are 1-kb genomic tiles that are significantly enriched for H3K27Ac over whole cell extract (WCE) and not enriched for other chromatin marks of higher priority. These regions are the ones displayed in the chromatin states maps that happen to fall into stitched H3K27Ac super-enhancers. For an extended H3K27Ac region to be classified as a super-enhancer, it must be enriched in H3K27Ac read density relative to all other H3K27Ac enhancer regions (not relative to WCE) for a given cell type.

TF enrichment analysis. We assessed the significance of overlap in TF binding and regions merged within super-enhancers by using the hypergeometric distribution. For each cell type, we only used TF peak regions in that cell type and super-enhancers as defined by chromatin data for that cell type. We used the same approach for measuring the TF binding enrichment at poised enhancers, or regions enriched for H3K4me1 and H3K27me3 histone modifications²⁹. For chromatin state transition analysis, we defined the initial state as ES cells and the next cellular state as dMS or one of the three germ layers (dEN, dME, and dEC).

We then carried out TF enrichment analysis using MNChIP binding data per cell type and different epigenetic state transitions into that cell type. P values were again calculated using the hypergeometric distribution, and were subsequently corrected for multiple hypothesis testing. This analysis was used for both chromatin state transitions and DNA methylation state transitions. For Fig. 6b, we identified all differentially methylated 1-kb tiles in the genome (mean methylation difference ≥ 0.15) between ES cells and the three germ layers. In addition, we also identified regions that transitioned from an HMR state to an H3K27Ac state, termed regions that lose methylation and gain H3K27Ac. We then carried out the enrichment analysis for TF binding in these regions as described above.

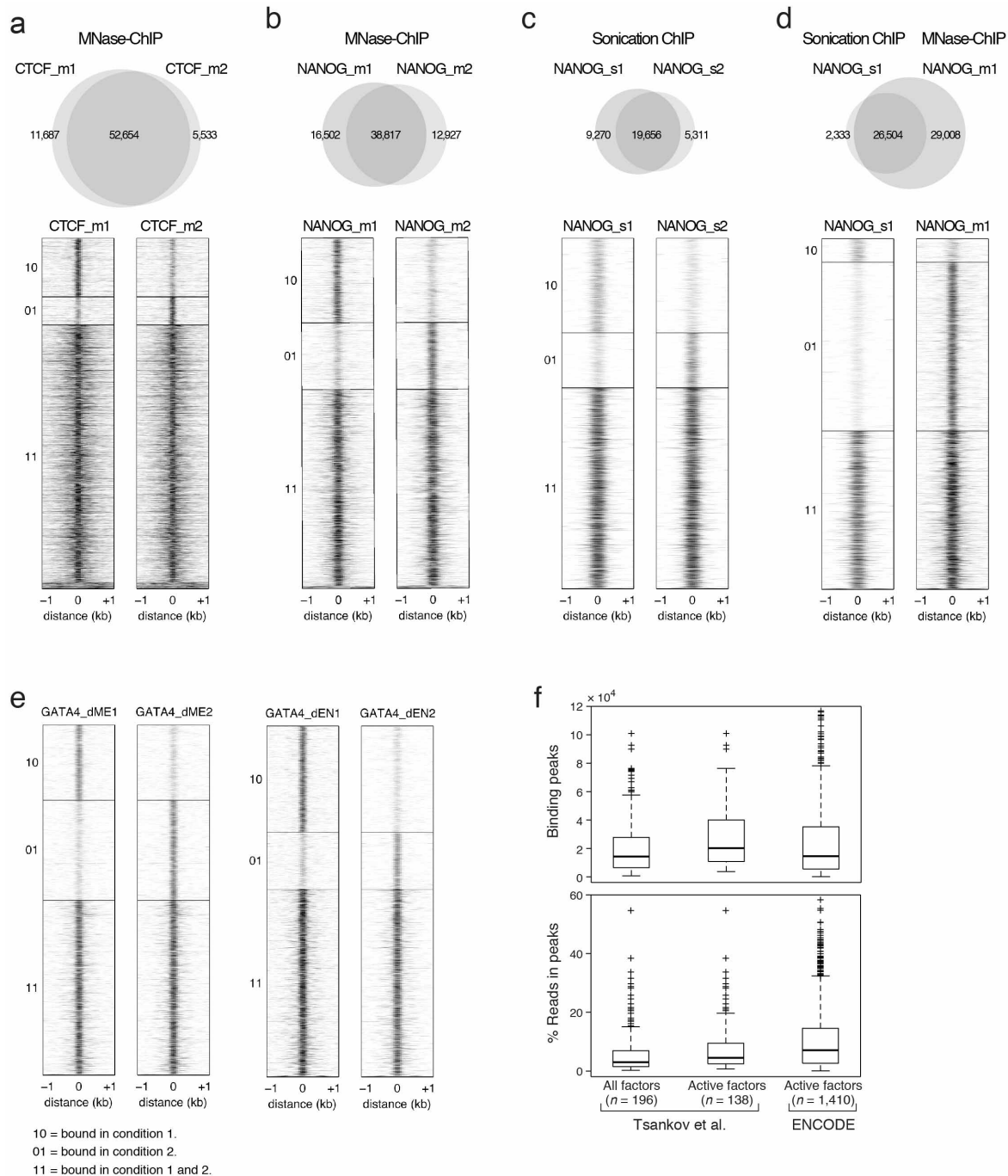
Heat maps and composite plots. Heat maps were generated for regions -1 kb to 1 kb from the centre of each merged TF peak, using bins of size 50 bp. ChIP occupancy was normalized to sequencing depth as described above. Binding events for two or three ChIP-seq experiments were merged before heat map generation using Homer, as described above. ChIP-seq composite plots were generated for regions -5 kb to 5 kb from the centre of each TF peak, using bins of size 200 bp. Signal was normalized to sequencing depth, where 1 represents the mean ChIP occupancy at regions furthest from the peaks. DNA methylation composite plots were generated for regions -2 kb to 2 kb from the centre of each TF peak, using bins of size 100 bp. Mean methylation was calculated by averaging of the methylation ratio at all unique CpGs within a given bin, excluding bins with no CpGs. P values for composite plots were calculated between two samples (for example, KD and control) by finding the normalized histone mark enrichment or normalized methylation level for each sample at 300 bp regions centred around each TF peak, and then using the paired t -test. Using region size of 1 kb or 600 bp led to the same biological conclusions. Reduced-representation bisulfite sequencing captured only 1,897 of the 42,477 GATA4 bound regions in dEN and 2,331 of 35,842 GATA4 bound regions in dME with sufficient CpG methylation coverage; hence only these regions were used for the composite plots in Fig. 6f, Extended Data Fig. 10e, and associated P value calculations.

34. Novershtern, N. *et al.* Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell* **144**, 296–309 (2011).

35. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

36. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).

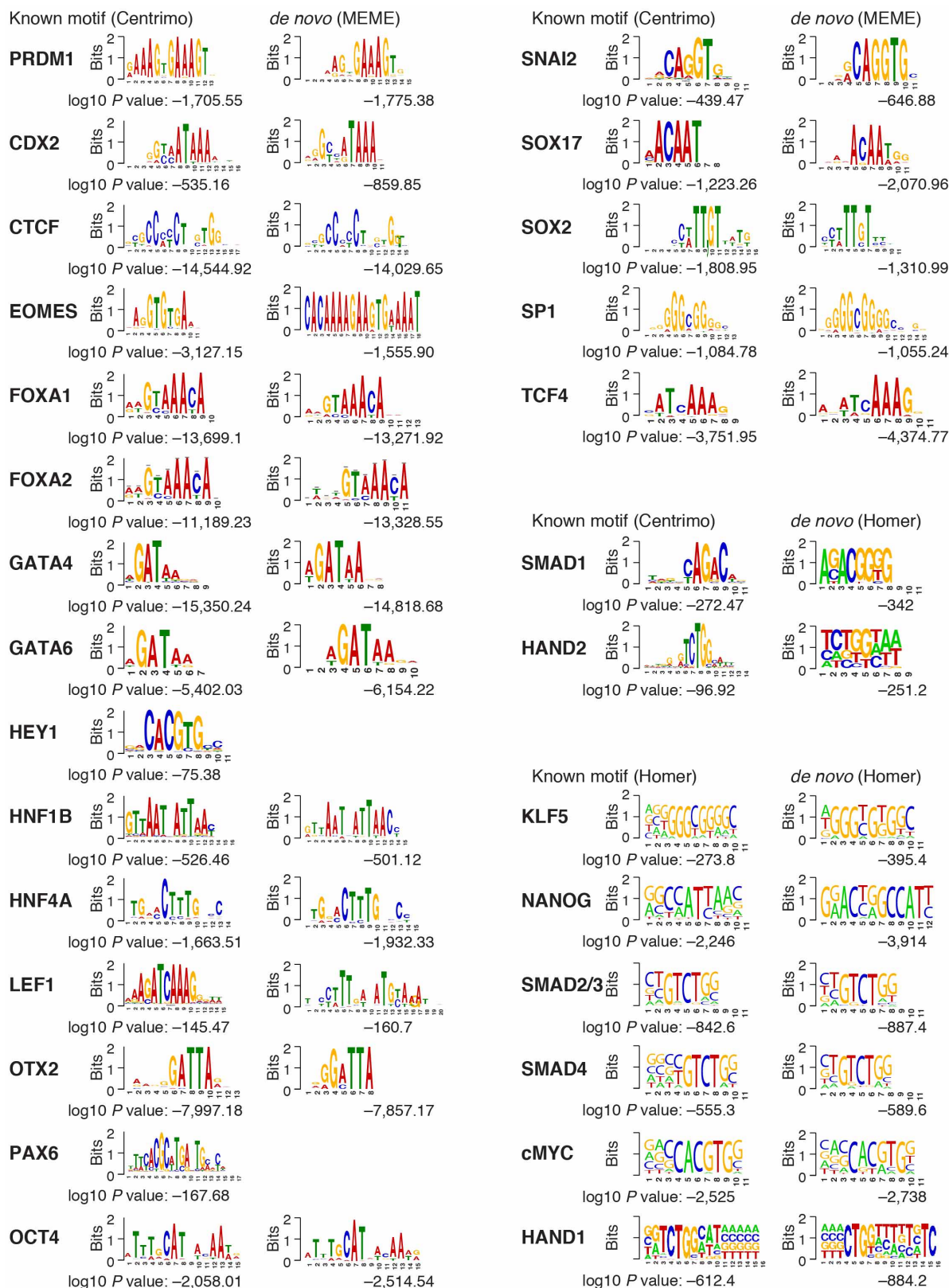
37. Pickrell, J. K., Gaffney, D. J., Gilad, Y. & Pritchard, J. K. False positive peaks in ChIP-seq and other sequencing-based functional assays caused by unannotated high copy number regions. *Bioinformatics* **27**, 2144–2146 (2011).
38. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
39. Bailey, T. L. & Machanick, P. Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Res.* **40**, e128 (2012).
40. Bailey, T. L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2**, 28–36 (1994).
41. Mikkelsen, T. S. *et al.* Comparative epigenomic analysis of murine and human adipogenesis. *Cell* **143**, 156–169 (2010).
42. Bao, Y., Vinciotti, V., Wit, E. & 't Hoen, P. A. C. Accounting for immunoprecipitation efficiencies in the statistical analysis of ChIP-seq data. *BMC Bioinformatics* **14**, 169 (2013).
43. Pauklin, S. & Vallier, L. The cell-cycle state of stem cells determines cell fate propensity. *Cell* **155**, 135–147 (2013).
44. Neph, S. *et al.* An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489**, 83–90 (2012).
45. You, J. S. *et al.* OCT4 establishes and maintains nucleosome-depleted regions that provide additional layers of epigenetic regulation of its target genes. *Proc. Natl Acad. Sci. USA* **108**, 14497–14502 (2011).
46. Mullen, A. C. *et al.* Master transcription factors determine cell-type-specific responses to TGF-beta signaling. *Cell* **147**, 565–576 (2011).
47. Morsli, H. *et al.* Otx1 and Otx2 activities are required for the normal development of the mouse inner ear. *Development* **126**, 2335–2343 (1999).
48. Greber, B. *et al.* FGF signalling inhibits neural induction in human embryonic stem cells. *EMBO J.* **30**, 4874–4884 (2011).
49. Pérez-Losada, J. *et al.* Zinc-finger transcription factor Slug contributes to the function of the stem cell factor c-kit signaling pathway. *Blood* **100**, 1274–1286 (2002).



Extended Data Figure 1 | MNase ChIP-seq (MNChIP-seq)

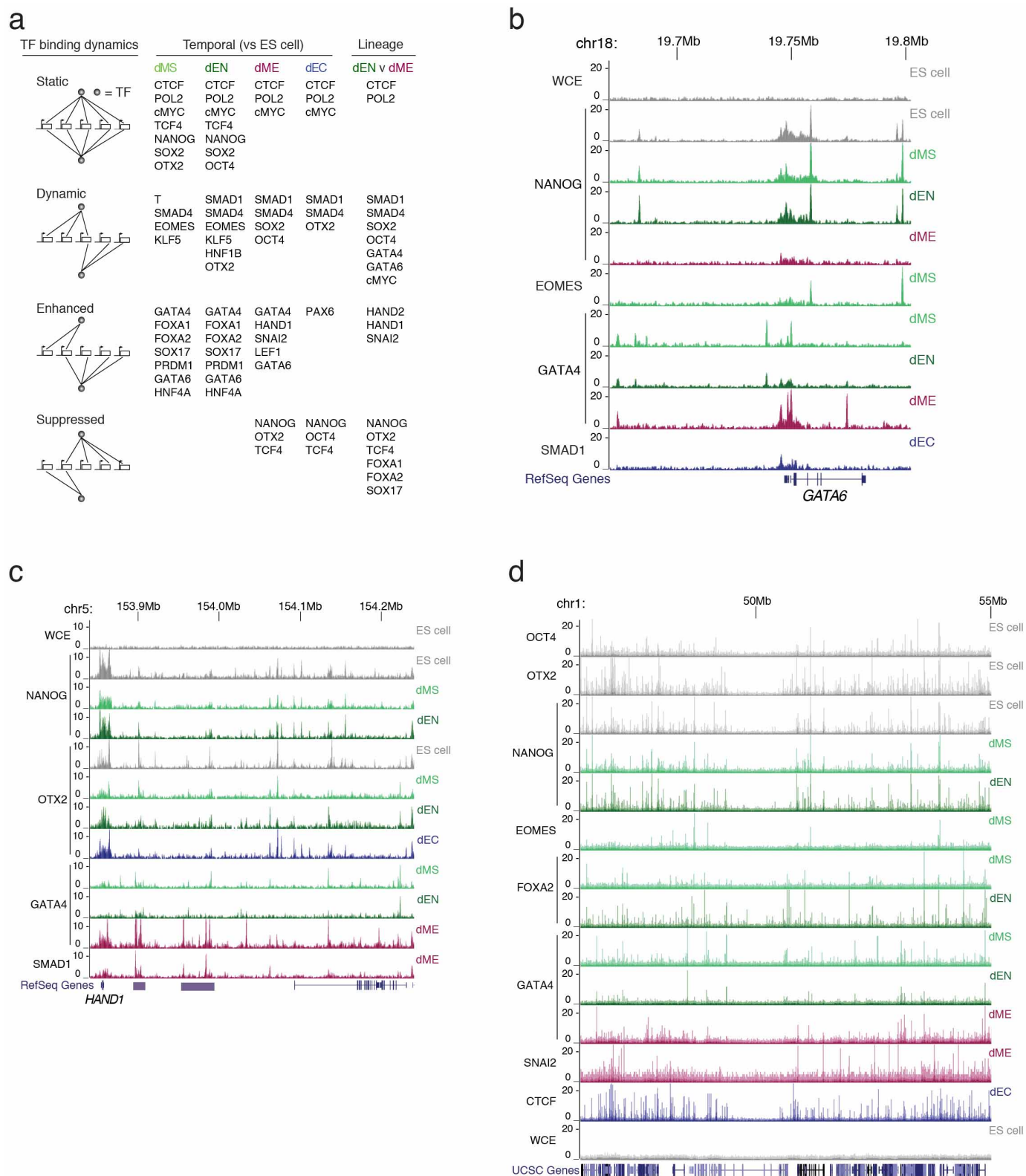
performance compared to sonication based ChIP-seq. **a**, Venn diagram (top) and corresponding heat maps (bottom) show high reproducibility of CTCF binding between biological replicates in ES cells using MNChIP-seq. Heat maps display normalized binding occupancy averaged using 50 bp bins. Regions are centred on the merged binding peaks for the two replicates, where 10 = regions bound in replicate 1, 01 = bound in replicate 2, and 11 = bound in both. **b**, Venn diagram (top) and corresponding heat maps (bottom) show high reproducibility of NANOG binding between biological replicates in ES cells using MNChIP-seq. Heat maps display normalized binding occupancy averaged using 50 bp bins. Regions are centred on the merged binding peaks for the two replicates, where 10 = regions bound in replicate 1, 01 = bound in replicate 2, 11 = bound in both. **c**, Venn diagram (top) and corresponding heat maps (bottom) show high reproducibility of NANOG binding between biological replicates in ES cells using sonication ChIP-seq. Reproducibility of NANOG binding using sonication based ChIP-seq is similar to reproducibility using MNChIP-seq. 10 = regions bound in replicate 1, 01 = bound in replicate 2, 11 = bound in both. **d**, Venn diagram (top) and corresponding heat

maps (bottom) show a higher sensitivity for capturing NANOG binding sites in ES cells using MNChIP-seq. Heat maps display normalized binding occupancy averaged using 50 bp bins. Regions are centred on the merged binding peaks for the two replicates, where 10 = regions bound in sonication ChIP-seq replicate, 01 = bound in MNChIP-seq replicate, 11 = bound in both replicates. **e**, Heat maps show high reproducibility of GATA4 binding in both dME (left) and dEN (right) using MNChIP-seq. Heat maps display normalized binding occupancy averaged using 50 bp bins. Regions are centred on the merged binding peaks for the two replicates, where 10 = regions bound in replicate 1, 01 = regions bound in replicate 2, 11 = regions bound in both. **f**, Top: number of significant binding peaks in our data set is comparable to that of 1,410 ENCODE TF ChIP-seq profiles (all currently available with matching peak and .bam files at UCSC). Bottom: the level of enrichment over background, as quantified by percentage of reads in peaks, is approximately 1.5 times less than that of the ENCODE TF binding data. ENCODE data was collected in cell types where the factors are known to be active; therefore, for this comparison we excluded all TF binding profiles from time points where the factors are not expressed and expected to be active (middle column).



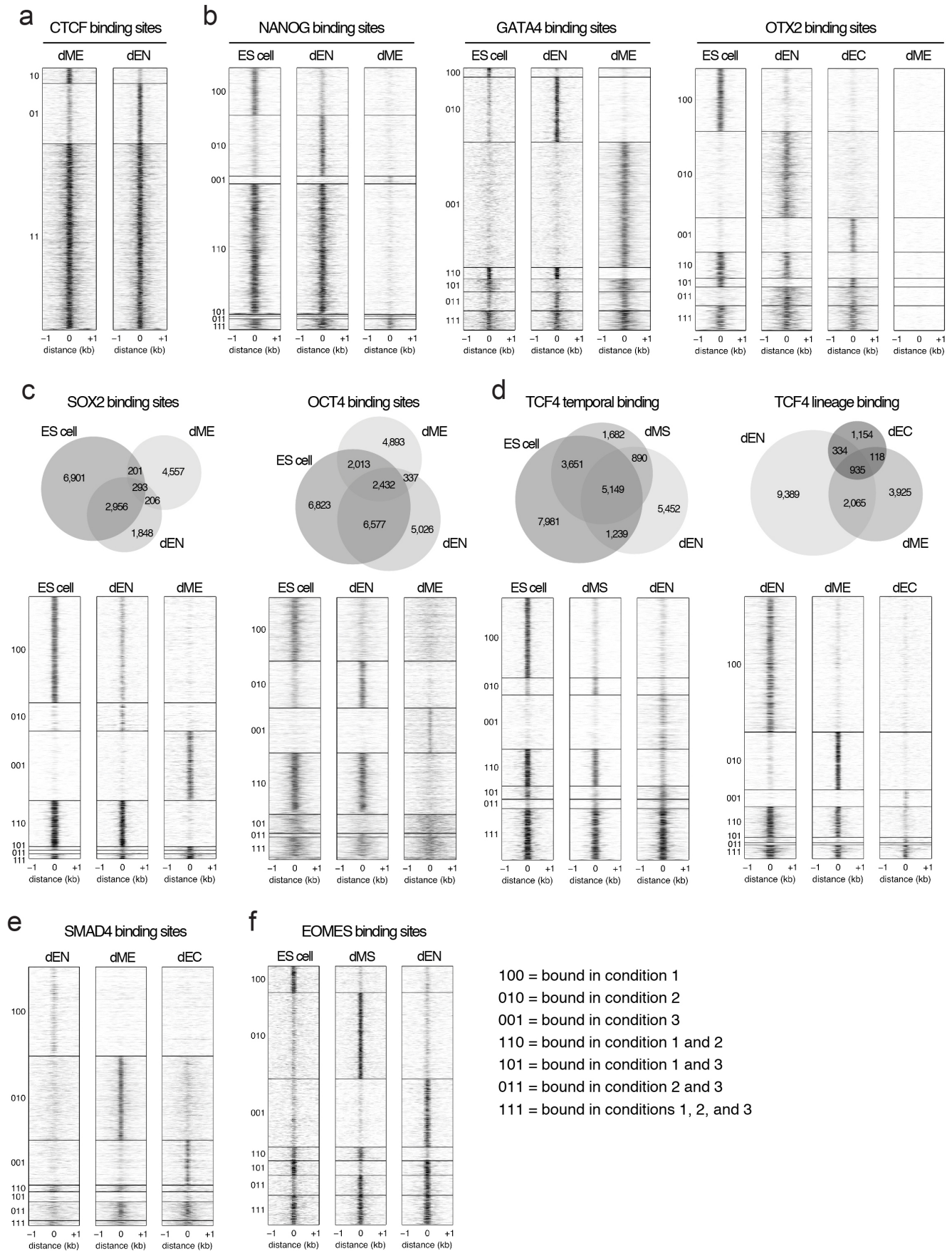
Extended Data Figure 2 | Motif analysis. 88% (28/32) of factors significantly associate with their known DNA binding motif ($P < 10^{-75}$). *De novo* motif discovery for these factors confirms their known motifs, which provides further validation for the antibody specificity. For SRF, REX1, STAT3, and TAL1 the motifs did not match the database motifs. To be conservative, we excluded

these factors from further analyses. For the remaining six factors, (POL2, SALL4, T, NR5A2, THAP11, TRIM28) we did not find a reliable DNA-binding motif in the database of 1,887 motifs combining TRANSFAC and Jolma *et al.* data sets²⁰.



Extended Data Figure 3 | Examples of TF binding dynamics across several loci. **a**, Binding dynamics for a number of selected TFs in the four differentiated cell types versus ES cells (temporal) and in dEN versus dME (cross-lineage). **b**, Normalized TF binding of NANOG, EOMES, GATA4, and SMAD1 shows distinct and germ layer specific regulation of the *GATA6* locus. **c**, Normalized TF binding at the *HAND1* locus shows very static binding for NANOG between cell types, somewhat dynamic binding of OTX2 in dEN and dEC, and more dynamic binding of GATA4 in dEN and dME. Purple boxes

upstream of *HAND1* mark long domains of H3K27Ac, which are highly enriched for GATA4 and SMAD1 binding in dME (bottom tracks). **d**, Normalized MNChIP-seq binding of multiple factors across different cell types show strong enrichments over whole cell extract (WCE) control (bottom track). The high similarity in CTCF binding between cell types suggests that chromatin loops, nuclear lamina interactions, and chromatin boundaries regulated by CTCF are largely preserved during early human ES cell differentiation.

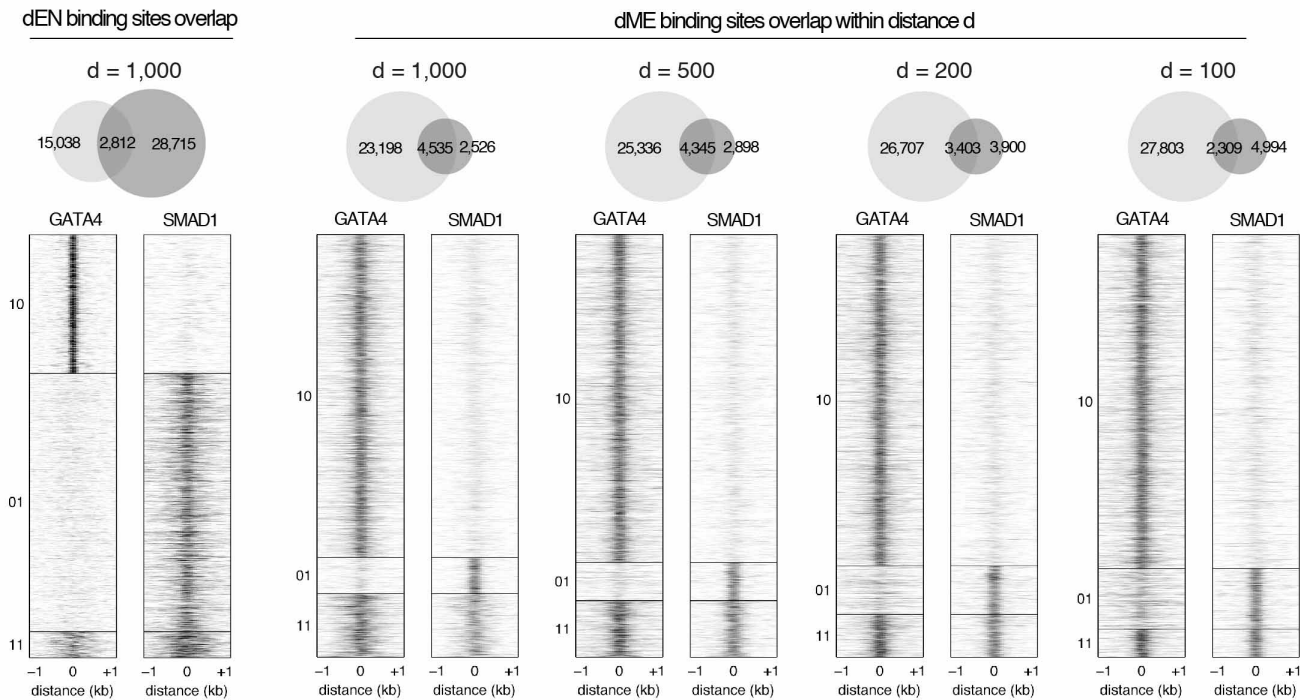


Extended Data Figure 4 | Venn diagrams and heat maps highlighting different TF binding dynamics in human ES cells and their derivatives.

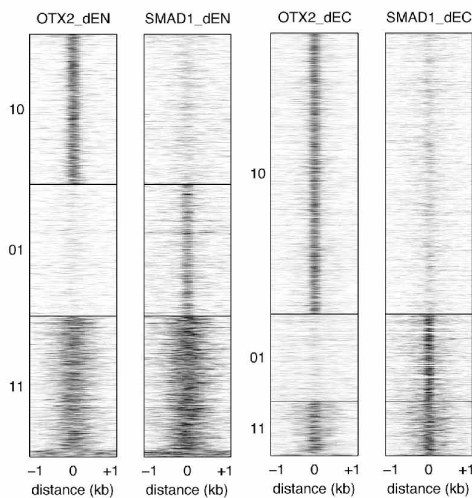
a, Heat maps show that CTCF binding overlaps highly in dEN and dME. Heat maps display normalized binding occupancy averaged using 50 bp bins. Regions are centred on the merged binding peaks for the two cell types, where 10 = regions bound in dME, 01 = bound in dEN, 11 = bound in both. **b**, Heat maps show that NANOG binding (left) is static in ES cells and dEN and suppressed in dME. In contrast, GATA4 binding (middle) is highly dynamic between dEN and dME and enhanced in the germ layers relative to ES cells. Finally, OTX2 binding (right) is dynamic in dEN and dEC relative to ES cells, but suppressed in dME. Heat maps display normalized binding occupancy averaged using 50 bp bins. Regions are centred on the merged binding peaks for the three conditions, where regions 100, 010, 001, 110, 101, 111 are defined in

legend on bottom right (panel **f**). **c**, Venn diagrams (top) and corresponding heat maps (bottom) show the binding dynamics of SOX2 (left) and OCT4 (right). Heat maps display normalized binding occupancy averaged using 50 bp bins. Regions are centred on the merged binding peaks for the three conditions, where regions 100, 010, 001, 110, 101, 111 are defined in legend in panel **f**. **d**, Venn diagram (top) and heat maps (bottom) show that TCF4 binding is temporally static in dMS and dEN (left) and suppressed in dME and dEC relative to dEN (right). **e**, Heat maps show that SMAD4 predominantly binds to unique regions in the three germ layers. **f**, Heat maps show that EOMES binding is enhanced from ES cells to dMS and dynamic in dEN. Heat maps display normalized binding occupancy averaged using 50 bp bins. Regions are centred on the merged binding peaks for the three conditions, where regions 100, 010, 001, 110, 101, 111 are defined in legend on the right.

a



b

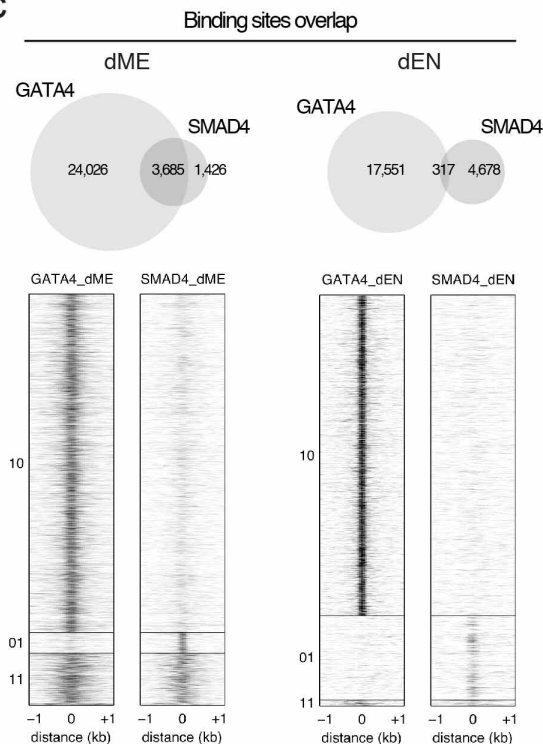


10 = bound in condition 1.

01 = bound in condition 2.

11 = bound in condition 1 and 2.

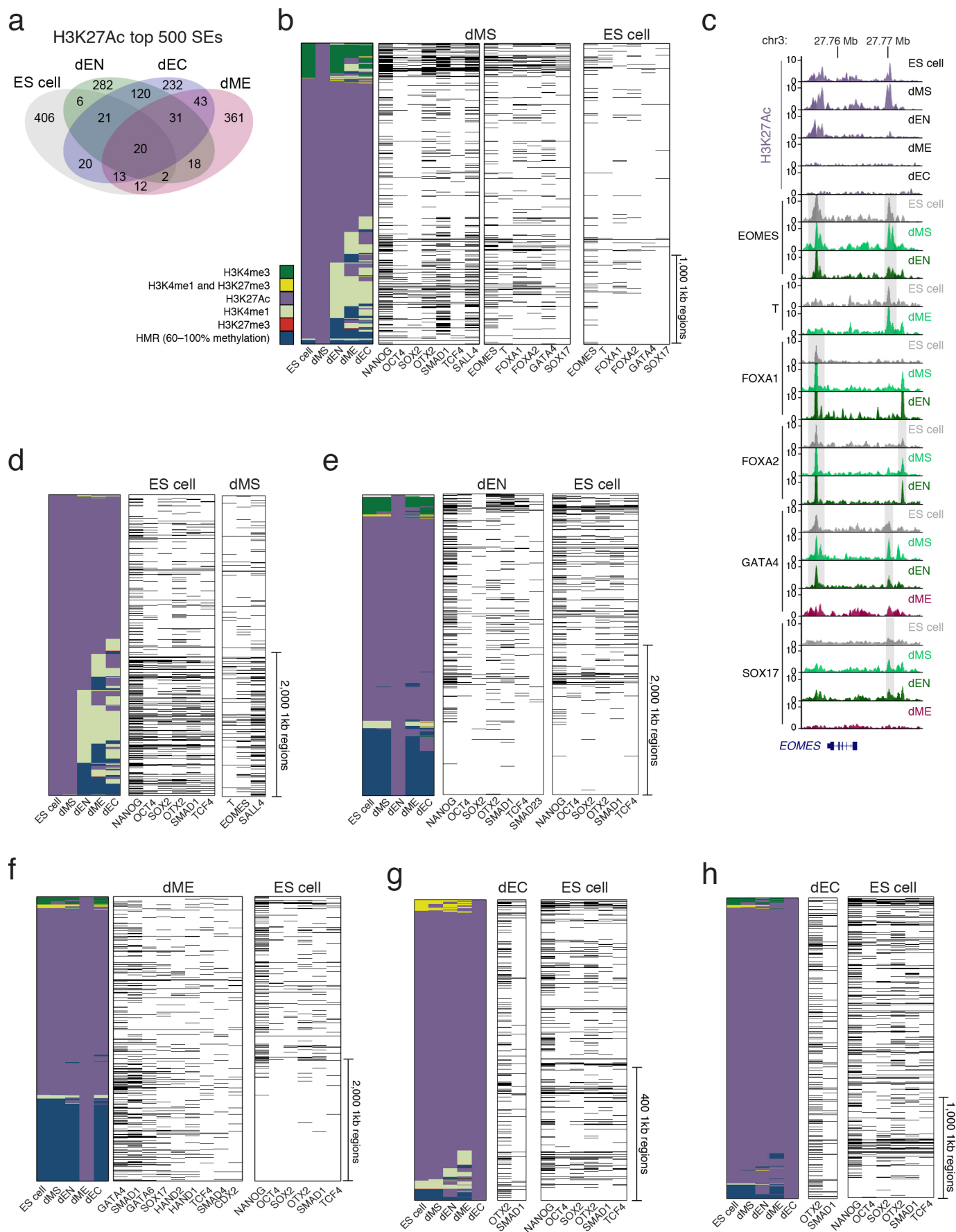
C



Extended Data Figure 5 | Heat maps of GATA4 and OTX2 co-binding relationship with SMAD1/4 in germ layers. a, Venn diagrams (top) and corresponding heat maps (bottom) show that overlap in binding between GATA4 and SMAD1 is smaller in dEN (left) than in dME (right). Heat maps display normalized binding occupancy averaged using 50 bp bins. Regions are centred on the merged binding peaks for the three conditions, where 10 = regions bound by factor 1, 01 = regions bound by factor 2, 11 = regions bound by both factors. Regions were considered co-bound if peaks for both factors occurred within distance d , set to 1000 bp for most analyses. Decreasing the distance d for dME to 500 bp has little effect. Setting d to 200 bp and 100 bp decreases co-bound peaks in dME by about 25% and 50%, respectively.

b, Heat maps show that overlap in binding between OTX2 and SMAD1 is higher in dEN (left) than in dEC (right). Heat maps display normalized binding occupancy averaged using 50 bp bins. Regions are centred on the merged binding peaks for the three conditions, where 10 = regions bound by factor 1, 01 = regions bound by factor 2, 11 = regions bound by both factors.

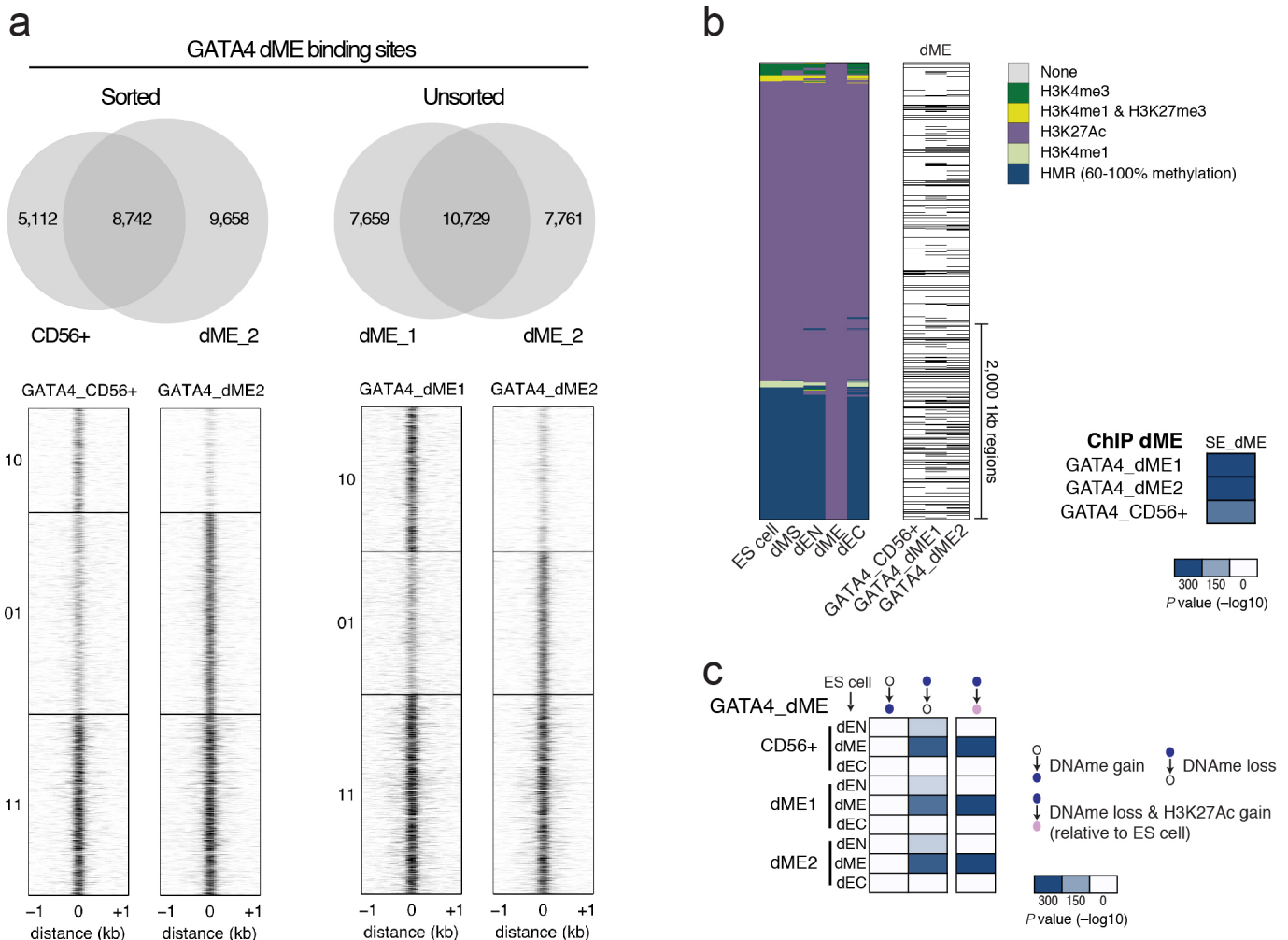
c, Venn diagrams (top) and corresponding heat maps (bottom) show that the overlap in binding between GATA4 and SMAD4 is greater in dME than in dEN. Heat maps display normalized binding occupancy averaged using 50 bp bins. Regions are centred on the merged binding peaks for the three conditions, where 10 = regions bound by factor 1, 01 = regions bound by factor 2, 11 = regions bound by both factors.



Extended Data Figure 6 | Extended H3K27Ac domains in the germ layers.

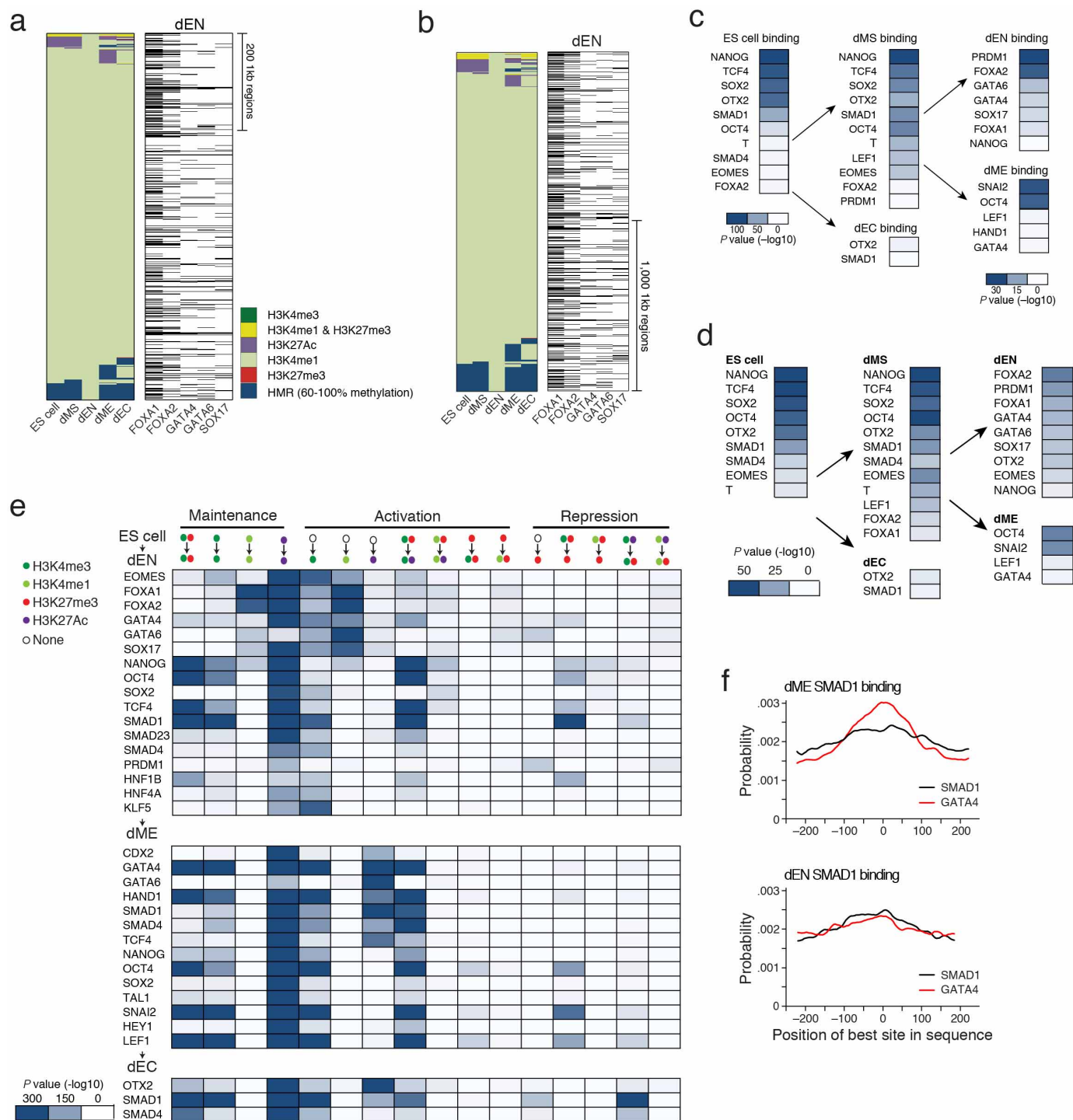
Related to Fig. 4. **a**, Venn diagram shows limited overlap between the top 500 extended H3K27Ac domains between cell types. **b**, Left, alternative lineage chromatin states of stitched dMS H3K27Ac super-enhancers ($n = 698$, merging 3,441 1-kb regions shown as rows in the heat map). Chromatin states (see Supplementary Information for detailed definitions of “extended H3K27Ac domains” and “H3K27Ac chromatin states”) that are displayed in the panel are defined in the legend (bottom left, HMR = highly methylated region). Centre, corresponding binding of the most enriched TFs in dMS. Black bars indicate TF binding. Right, corresponding binding of selected factors in ES cells. **c**, Genome browser tracks for H3K27Ac across all cell types and normalized TF binding in selected cell types for EOMES, T, FOXA1/2, GATA4, and SOX17 over the *EOMES* locus. Grey bars highlight regions where TF binding is present in ES cells and at later stages in differentiation, suggesting that these loci are primed for binding by these factors in ES cells. Although we cannot distinguish whether this happens in all cells or just a subpopulation, it is tempting to speculate that this binding occurs in the subset of cells in G1, which is the population that is most responsive to differentiation cues⁴³. This would also be in line with DNase I footprint studies that reported usage of EOMES DNA-binding sites in human ES cells⁴⁴. **d**, Left, alternative lineage chromatin states of stitched H3K27Ac super-enhancers in ES cells ($n = 1,052$, merging 4,191 1-kb regions shown as rows in the heat map). Chromatin states that are displayed in the panel are explained in the legend in panel **b** (bottom left, HMR = highly methylated region). Centre, corresponding binding of the most enriched TFs in ES cells. Black bars indicate TF binding. OSN and OTX2 are the most enriched factors. Interestingly, OTX2 was recently shown to play an important role in the mouse naive to primed pluripotent state transition, a cellular state considered to be similar to human ES cells²⁶. Right, corresponding binding data for T, EOMES, and SALL4 in dMS, showing that these key dMS regulators are present at many of these super-enhancers in the next stage of differentiation. **e**, Left, alternative lineage chromatin states of stitched H3K27Ac super-enhancers in dEN ($n = 1,152$, merging 4,051 1-kb regions shown as rows in the heatmap). States are defined as in panel **b**. Centre, corresponding binding of the most enriched TFs in dEN. Black bars indicate TF binding. Right, corresponding binding of selected TFs in ES cells shows that these factors occupy many of these regions in the undifferentiated state. Despite the fact that H3K27Ac domains are highly unique in the different cell types, we note that OSN, OTX2 and SMAD1 binding in undifferentiated ES cells is observed before the other factors that will mediate the transition to super-enhancer status in the three germ layers (**e–h**, right panels). Similarly, as noted above, regulators of super-enhancers in the germ layers also associate with these regions already in the pluripotent state. This might suggest that TF binding at germ layer specific H3K27Ac domains in the ES cells could be involved or necessary for the future handoff. Possible roles could include active

regulatory binding or a way to simply mark super-enhancers; alternatively, it could also provide an active protection from silencing by the highly expressed DNA methylation machinery. In this context it is worth noting that OSN binding in the undifferentiated cells is depleted in a subset of super-enhancers that are highly methylated (**e–h**, bottom right) suggesting a possible binding sensitivity to DNA methylation, which has been reported for OCT4 (ref. 45). **f**, Left, alternative lineage chromatin states of dME H3K27Ac super-enhancers ($n = 1,129$, merging 4,717 1-kb regions shown as rows in the heat map). States are defined as in panel **b**. Centre, corresponding binding of the most enriched TFs in dME. GATA4 and SMAD1 are the most highly enriched factors at dME super-enhancers. Globally, GATA4 also interacts significantly with SMAD1 and SMAD4 in dME (hypergeometric $P < 10^{-300}$) but less so in dEN (Fig. 3a, Extended Data Fig. 5c). This suggests that GATA4 interacts with SMAD1/4 at genomic targets and specifically at super-enhancers to act as a possible key regulator of the transition from pluripotent to a mesodermal state in response to BMP signalling. Recent studies have reported that master regulators in various cell types interact with TFs downstream of key signalling pathways in a similar manner⁴⁶. Black bars indicate TF binding. Right, corresponding binding of selected factors in ES cells. **g**, Left, alternative lineage chromatin states of stitched H3K27Ac super-enhancers in dEC ($n = 506$, merging 908 1-kb regions shown as rows in the heat map). States are defined as in panel **b**. Centre, corresponding binding of the most enriched TFs in dEC. Black bars indicate TF binding. Right, corresponding binding of selected TFs in ES cells shows that these factors occupy many of these regions in the undifferentiated state. OTX2 is known to play important roles in brain, craniofacial, and sensory organ development^{28,47,48}. In mice, *Otx2* is required from E10.5 onward to regulate neuronal subtype identity and neurogenesis in the midbrain²⁸, and inhibition of FGF signalling upregulates OTX2 and subsequently induces the neuroectodermal regulator PAX6 (ref. 48). Complementing these previous studies, our results suggest that it may play a central role in mediating the transition from pluripotency to early ectoderm. Interestingly, in dEC OTX2 does not globally associate with SMAD1 outside of super-enhancers to the same degree as in dEN (Fig. 3a). Taken together, we observe differential co-binding between SMAD1 and GATA4 or OTX2 in the respective germ layers that is linked to differential signalling, which may guide the remodelling of the associated chromatin. **h**, Left, alternative lineage chromatin states of the top 3,000, 1-kb-long H3K27Ac enhancers in dEC, showing a comparable number of genomic regions as in the other cell types. States are defined as in panel **b**. Centre, corresponding binding of OTX2 and SMAD1 in dEC shows a higher enrichment for these factors at H3K27Ac enhancers than when only surveying the top 908 1-kb regions (panel **d**). Black bars indicate TF binding. Right, corresponding binding of selected TFs in ES cells shows that these factors occupy many of these regions in the undifferentiated state.



Extended Data Figure 7 | Quantification of cell sorting in dME on GATA4 binding and enrichment analysis. **a**, Left, Venn diagrams (top) and heat maps (bottom) show that the overlap in binding between GATA4 ChIP-seq in sorted CD56⁺ cells and unsorted dME cells is very similar. In particular, the unique binding sites in unsorted cells (*y* axis label 01) also show visible but less significant binding in sorted cells, arguing that unsorted cells do not add many false positive peaks. Conversely, unique binding sites in sorted cells (*y* axis label 10) show that less than half of these sites are truly unique, or with no detectable binding in unsorted cells. Right, Venn diagrams (top) and heat maps (bottom) shows the overlap in binding between two GATA4 ChIP-seq replicates in unsorted dME populations. The overlap in binding between sorted and unsorted cells shown on the left. **b**, Left, alternative lineage chromatin states of dME H3K27Ac super-enhancers ($n = 1,129$, merging 4,717 1-kb regions shown as rows in the heat map). States are defined in legend (top right, HMR = highly methylated region). Centre, corresponding binding of GATA4 in sorted CD56⁺ cells, and two unsorted dME replicates (dME1 and

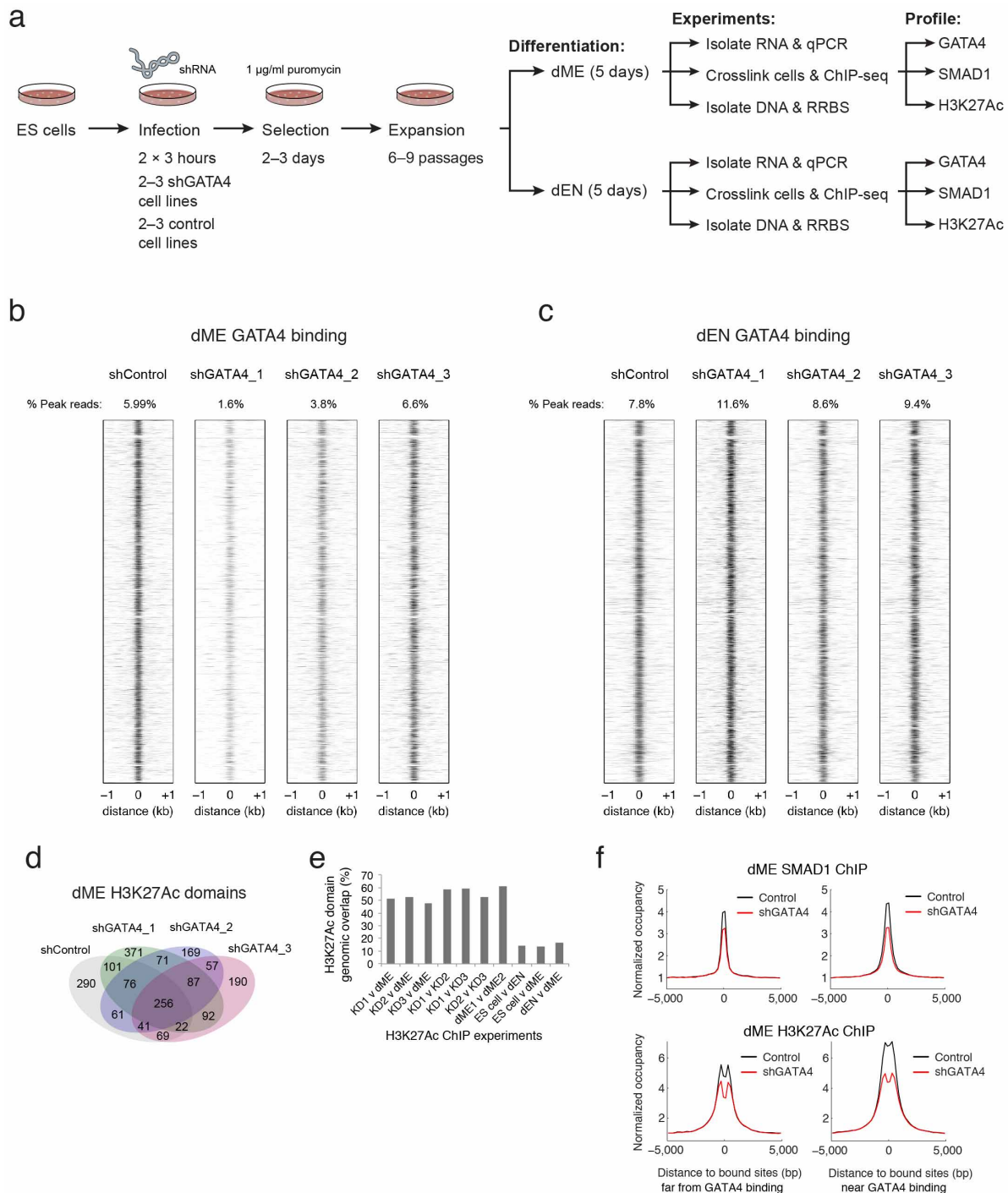
dME2). Black bars indicate TF binding. Right, enrichment *P* values ($-\log_{10}$) for GATA4 binding at H3K27Ac super-enhancers are slightly more significant (hypergeometric $P < 10^{-300}$) for unsorted cells than for sorted cells (hypergeometric $P < 10^{-225}$). This shows that the conclusions for GATA4 in dME are largely unaffected by cell sorting. Moreover, since our enrichment analysis compares overlaps of binding at thousands of sites, this comparison argues that the analysis is in general robust to using unsorted cell populations. **c**, Enrichment *P* values ($-\log_{10}$) for the overlap in TF binding and regions that gain or lose DNA methylation relative to ES cells (see Supplementary Information). Possible transition states are defined at the top. Heat maps display the enrichment of GATA4 binding in sorted CD56⁺ cells, and two unsorted dME replicates (dME1 and dME2). Unsorted cells have similar enrichment *P* values (hypergeometric $P < 10^{-300}$) to sorted cells (hypergeometric $P < 10^{-300}$). This shows that the methylation conclusions for GATA4 in dME are largely unaffected by cell sorting and again argues that our enrichment analysis is robust to using unsorted cell populations.



Extended Data Figure 8 | Regulation of poised enhancers and other epigenetic state transitions.

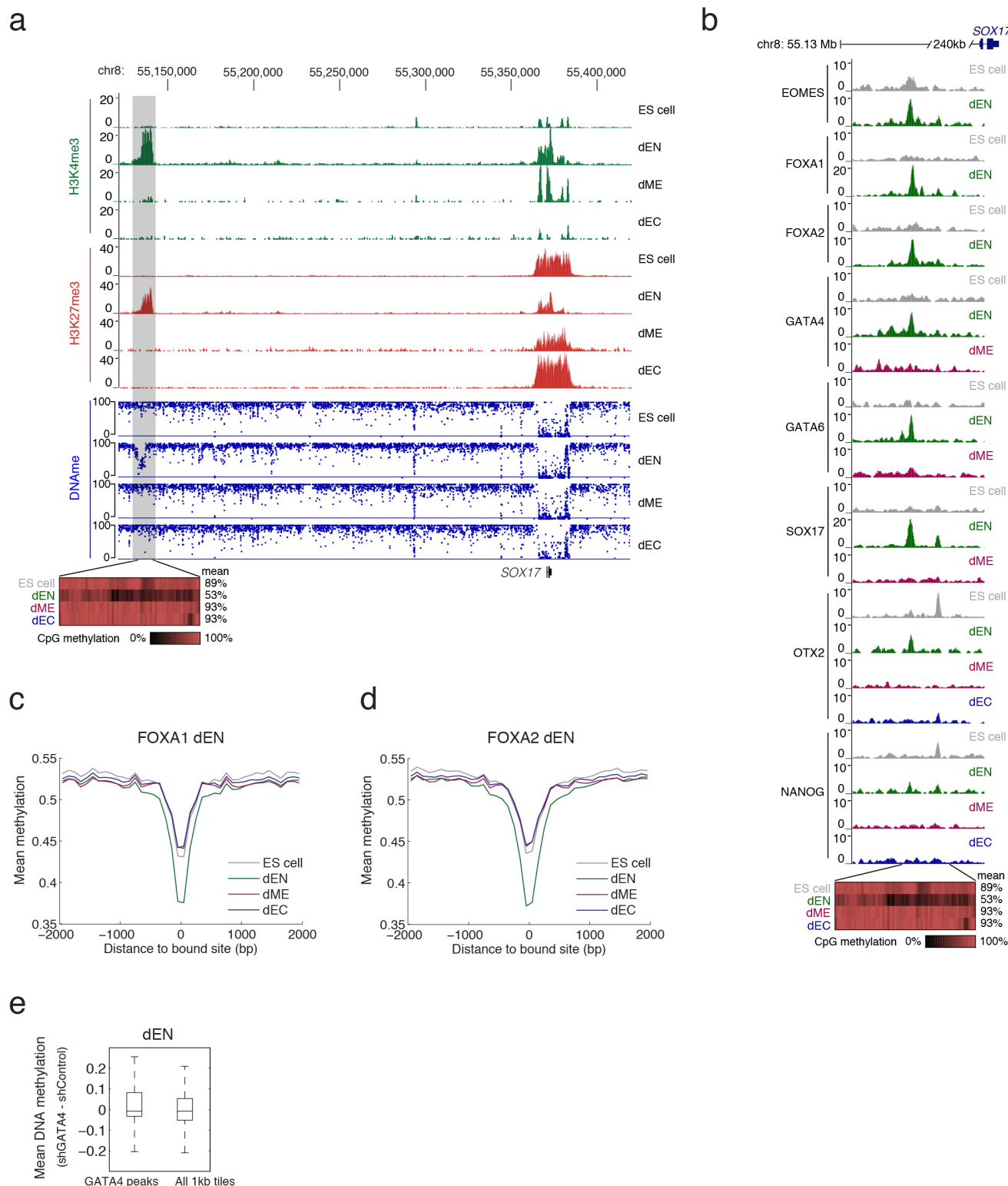
Related to Fig. 5. **a**, Left, alternative lineage chromatin states of dEN H3K4me1 super-enhancers ($n = 309$, merging 760 1-kb regions shown as rows in the heat map). Chromatin states that are displayed in the panel are explained in the legend (bottom right, HMR = highly methylated region). Right, corresponding binding of the most enriched TFs in dEN, where the black bars indicate TF binding. **b**, Left, alternative lineage chromatin states of the top 2,000 1-kb-long dEN H3K4me1 enhancers in dEN (shown as rows in the heat map). Chromatin states that are displayed in the panel are defined in the legend in panel **a** (bottom right, HMR = highly methylated region). Right, corresponding binding of the most enriched TFs in dEN, where the black bars indicate TF binding. Increasing the number of regions displayed shows a higher enrichment for dEN factors at H3K4me1 enhancers than when only surveying the top 760 1-kb regions in panel **a**. **c**, Hypergeometric P values ($-\log_{10}$) for the most significant overlaps between all poised putative enhancers (H3K27me3 and H3K4me1) and each TF's binding profile in the respective cell type. Enrichment P values for dEN and dME (right column) are lower than in ES cells, which is likely the result of the overall smaller number of poised enhancers in those two germ layers. The scale is therefore adjusted for dEN and dME as shown in the respective P value legends. In ES cells, we find that poised enhancers are highly enriched for binding by OSN, OTX2, TCF4 and SMAD1 in the pluripotent state (**c**, **d**). In dMS, we see the same regulators along with T, EOMES, and LEF1 are present at poised enhancers (**c**, **d**, centre). In contrast, poised enhancers in dEN show strong enrichment for PRDM1 and many of the regulators mentioned above (**c**, **d**, right). Lastly, in dME, we find enrichment for SNAI2, which is known for

its activity in mesoderm including blood development⁴⁹. **d**, Summary table of hypergeometric P values ($-\log_{10}$) displaying the most significant overlaps between the top 500 poised enhancers (H2K27me3 and H3K4me1) and each TF's binding profile within a given cell type (ES cells, left; dMS, centre; dEC, bottom centre; dEN and dME, right). Enrichment P values are more comparable between ES cells and the germ layers, since we compare TF binding with the same number of poised enhancers (500) in each cell type. The results are consistent with **c**, showing that the same factors are most enriched as when comparing to all poised enhancers. **e**, Table of hypergeometric P values ($-\log_{10}$) in overlap between TF binding and regions with different chromatin state transitions (relative to ES cells) within each germ layer (dEN, top; dME, middle; dEC, bottom; see Supplementary Information). Possible epigenetic state transitions are shown on top and states are defined in legend on the top left. Globally, we find a much stronger enrichment for gain of H3K4me1 in dEN than in dME, particularly for the endoderm factors present at the most methylated H3K4me1 domains. Conversely, in dME we find a strong association between remodelling of H3K27Ac and the dME factors that reside at H3K27Ac genomic regions. In concordance with this global trend, GATA4 is associated with dynamics of H3K4me1 in dEN and H3K27Ac in dME. **f**, Probability (y axis) of the best match to a given motif (SMAD1 and GATA4) occurring at a given position at regions centred on SMAD1 binding in dME (top) and dEN (bottom). This probability is based only on regions that contain at least one match with score greater than the minimum score defined for this motif by the default settings in Centrimo³⁹. The position of the best GATA4 DNA binding sites (red) are more centrally enriched ($P < 10^{-241}$, Centrimo³⁹) at SMAD1 ChIP-seq peaks in dME (top) than in dEN (bottom).



Extended Data Figure 9 | GATA4 knockdown experiments in dEN and dME. **a**, Experimental design and data collected for the GATA4 knockdown (KD) and control experiments in dEN and dME (see Supplementary Information for details). **b**, Heat maps of GATA4 normalized occupancy at GATA4 targets (columns) in control and KD cell lines at corresponding genomic regions. GATA4 occupies very similar loci in control and KD cell lines in dME. **c**, Heat maps of GATA4 normalized occupancy at GATA4 targets (columns) in control and KD cell lines at corresponding genomic regions. GATA4 occupies very similar loci in control and KD cell lines in dEN. **d**, Venn diagram of dME H3K27Ac super-enhancers detected using H3K27Ac data in shControl and 3 shGATA4 KD cell lines. Super-enhancers in the shGATA4 KD lines 1, 2, and 3 overlap with super-enhancers in the shControl cell line at a much higher rate than different cell types in Fig. 4b. **e**, Pairwise rate of overlap between super-enhancers detected using different H3K27Ac ChIP experiments. Super-enhancers in the shGATA4 KD lines 1, 2, and 3 overlap with super-enhancers in the shControl cell line at a rate of 51.6%, 52.7%, and

47.4% (left-most bars). In comparison, the KD replicates overlap with one another at a rate of 58.8%, 59.3%, and 52.3%, and wild-type dME replicates overlap at a rate of 61.2% (middle bars). This shows that the GATA4 KD does not affect the genomic location of dME H3K27Ac domains greatly, as the overlap in domains between control and KD cell lines is only slightly lower than between H3K27Ac ChIP-seq replicates in dME. In contrast, the number of super-enhancers in common between different cell types ES cell, dEN, and dME is much lower at 14.3%, 13.7%, and 16.7% (right-most bars). Percentages are calculated relative to the experiment with fewer super-enhancers detected. **f**, Normalized SMAD1 (top) and H3K27Ac (bottom) mean occupancy is lower in dME for the shRNA KD lines versus control lines at SMAD1 sites both far from (distance > 1 kb, left panel) and near (distance ≤ 1 kb, right panel) from GATA4 binding (see Supplementary Information for details). The smaller decrease in occupancy away from GATA4 binding may be due to indirect effects, such as lower SMAD1 expression or co-binding with other unknown TFs.



Extended Data Figure 10 | TF binding associates with specific loss of DNA methylation in dEN. Related to Fig. 6. **a**, Top, genome browser tracks for H3K4me3 and H3K27me3 across four of the cell types over the *SOX17* locus, zooming out on the region shown in Fig. 6a. Bottom, whole genome bisulfite sequencing (WGBS)-based CpG methylation measurements. Specific loss of DNA methylation in dEN and associated chromatin remodelling to a poised state (H3K4me3 and H3K27me3) occurs 240 kb upstream of *SOX17*, which coincides with loss of H3K27me3 and gain of H3K4me3 mark near the *SOX17* gene. **b**, Top, genome browser tracks for selected TFs in different cell types upstream of *SOX17*. Bottom, WGBS-based CpG methylation measurements, where each rectangle represents a single CpG. Specific loss of DNA methylation

in dEN coincides with specific binding of several endoderm factors. OTX2 and NANOG also bind nearby this region in ES cells. **c**, WGBS-based average CpG methylation level of 100-bp tiles over FOXA1 bound dEN targets in ES cells and the three germ layers shows a specific depletion of DNA methylation in dEN. **d**, WGBS-based average CpG methylation level of 100-bp tiles over FOXA2 bound dEN targets in ES cells and the three germ layers shows a specific depletion of DNA methylation in dEN. **e**, Distributions of mean DNA methylation difference in dEN between GATA4 KD and control cell lines at 1-kb regions centred on dEN GATA4 targets (left, $P < 10^{-10}$, paired *t*-test) and at all 1-kb regions in the genome (right, $P = 1$, paired *t*-test).

Integrative analysis of haplotype-resolved epigenomes across human tissues

Danny Leung^{1*}, Inkyung Jung^{1*}, Nisha Rajagopal^{1*}, Anthony Schmitt¹, Siddarth Selvaraj¹, Ah Young Lee¹, Chia-An Yen¹, Shin Lin^{2,3}, Ying Lin^{2,4}, Yunjiang Qiu¹, Wei Xie⁵, Feng Yue⁶, Manoj Hariharan⁷, Pradipta Ray⁸, Samantha Kuan¹, Lee Edsall¹, Hongbo Yang⁹, Neil C. Chi^{9,10}, Michael Q. Zhang^{8,11}, Joseph R. Ecker⁷ & Bing Ren^{1,10,12,13}

Allelic differences between the two homologous chromosomes can affect the propensity of inheritance in humans; however, the extent of such differences in the human genome has yet to be fully explored. Here we delineate allelic chromatin modifications and transcriptomes among a broad set of human tissues, enabled by a chromosome-spanning haplotype reconstruction strategy¹. The resulting large collection of haplotype-resolved epigenomic maps reveals extensive allelic biases in both chromatin state and transcription, which show considerable variation across tissues and between individuals, and allow us to investigate *cis*-regulatory relationships between genes and their control sequences. Analyses of histone modification maps also uncover intriguing characteristics of *cis*-regulatory elements and tissue-restricted activities of repetitive elements. The rich data sets described here will enhance our understanding of the mechanisms by which *cis*-regulatory elements control gene expression programs.

We performed chromatin immunoprecipitation followed by sequencing (ChIP-seq) to generate extensive data sets profiling 6 histone modifications across 16 human tissue types from four individual donors (181 data sets). In combination with previously published data sets^{2,3}, we conducted in-depth analyses across 28 cell/tissue types, covering a wide spectrum of developmental states, including embryonic stem cells, early embryonic lineages and somatic primary tissue types representing all three germ layers (Fig. 1a) (protocols received approval from IRB/ESCRO and Mid-American Transplant Services, and research consent was obtained from families). The modifications demarcate active promoters (histone H3 lysine 4 trimethylation (H3K4me3) and H3 lysine 27 acetylation (H3K27ac)), active enhancers (H3 lysine 4 monomethylation (H3K4me1) and H3K27ac), transcribed gene bodies (H3 lysine 36 trimethylation (H3K36me3)) and silenced regions (H3K27 or H3 lysine 9 trimethylation (H3K27me3 and H3K9me3, respectively))^{4,5}. We systematically identified *cis*-regulatory elements by employing a random-forest-based algorithm (RFECs)^{2,6}, predicting a total of 292,495 enhancers (consisting of 175,912 strong enhancers with high H3K27ac enrichment) across representative samples of all 28 tissues types (Supplementary Table 1). We additionally identified 24,462 highly active promoters with strong H3K4me3 enrichment (see Supplementary Table 2). Subsequently, we defined tissue-restricted promoters ($n = 10,396$) and enhancers ($n = 115,222$) (Extended Data Fig. 1a). Consistent with previous studies^{7–9}, enhancers appear more tissue-restricted than promoters and cluster along developmental lineages (Extended Data Fig. 1a, b). Moreover, tissue-restricted enhancers were enriched for putative binding motifs of particular transcription factors known to be important



in maintaining the cell/tissue type's identity and function^{10–15} (Extended Data Fig. 2).

Recent studies showed

that particular repetitive elements, such as endogenous retroviruses (ERVs), could participate in transcriptional regulation during mammalian development^{16–18}. Given the representation of samples available, we systematically examined histone modifications at different classes of ERVs. While most are inactive, subsets, especially class I ERVs (ERV-I), are marked by H3K27ac in a tissue-restricted manner (Extended Data Fig. 3a and b). For instance, HERV-H element activities are restricted to human embryonic stem cells (hESCs) (Extended Data Fig. 3c, d). Furthermore, some ERVs carried marks of active promoters or enhancers (Extended Data Fig. 3d, e). We also observed that the LTR12C subfamily had substantial H3K27ac enrichment across different tissues (Extended Data Fig. 3e, f). Notably, the individual members appeared to be tissue restricted, suggesting that although the subfamily can be classified as non-tissue restrictively active, individual LTR12C elements were active only in distinct tissue/cell types (Extended Data Fig. 3e). Taken together, the data illustrate that human ERVs display precisely controlled patterns of activity in distinct tissues.

Intriguingly, 15.2% ($n = 3,717$) of strong promoters were also predicted as enhancers in other tissues, analogous to observations in mice, where intragenic enhancers act as promoters to produce cell-type-specific transcripts¹⁹. These sites possessed histone modification signatures of active enhancers in some tissue/cell types but were enriched with active promoter marks in others. We termed these sequences *cis*-regulatory elements with dynamic signatures (cREDS). For example, 1,321 cREDS enhancers showed enrichment of H3K27ac and H3K4me1 and a striking depletion of H3K4me3 in lung (Fig. 1b, c and Supplementary Table 3). However, the signature shifted to that of active promoters in other tissues (Fig. 1b, c). cREDS are also found in other cell/tissue types (Extended Data Fig. 4a). To determine whether cREDS have dual functions, we selected a subset of promoter-marked elements and validated their function with a luciferase reporter assay in hESCs. The majority (7 out of 10) showed promoter activity (Extended Data Fig. 4b). Similarly, 10 of 11 selected cREDS with enhancer signatures in hESCs also functioned as enhancers (Extended Data Fig. 4c). Additionally, subsets of enhancers previously validated in transgenic mice also possessed dynamic signatures (Extended Data Fig. 5)²⁰. Furthermore, we selected two cREDS, predicted as enhancers in the left heart ventricle, with significant cap

¹Ludwig Institute for Cancer Research, La Jolla, California 92093, USA. ²Department of Genetics, Stanford University, 300 Pasteur Drive, M-344 Stanford, California 94305, USA. ³Department of Cardiovascular Medicine, Stanford University, Falk Building, 870 Quarry Road Stanford, California 94304, USA. ⁴Department of Surgery, Washington University School of Medicine, 660 S. Euclid Ave, Campus Box 8109, St Louis, Missouri 63110, USA. ⁵Tsinghua University–Peking University Center for Life Sciences, School of Life Sciences, Tsinghua University, Beijing 100084, China. ⁶Department of Biochemistry and Molecular Biology, College of Medicine, The Pennsylvania State University, Hershey, Pennsylvania 17033, USA. ⁷Genomic Analysis Laboratory, Howard Hughes Medical Institute, The Salk Institute for Biological Studies, La Jolla, California 92093, USA. ⁸Biological Sciences, Center for Systems Biology, The University of Texas at Dallas, Richardson, Texas 75080, USA. ⁹Department of Medicine, Division of Cardiology, University of California, San Diego, California 92093-0613, USA. ¹⁰Institute of Genomic Medicine, University of California, San Diego, California 92093, USA. ¹¹Bioinformatics Division, Center for Synthetic and Systems Biology, TNLIST Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing 100084, China. ¹²Department of Cellular and Molecular Medicine, University of California San Diego, La Jolla, California 92093, USA. ¹³UCSD Moores Cancer Center, University of California San Diego, La Jolla, California 92093, USA.

*These authors contributed equally to this work.

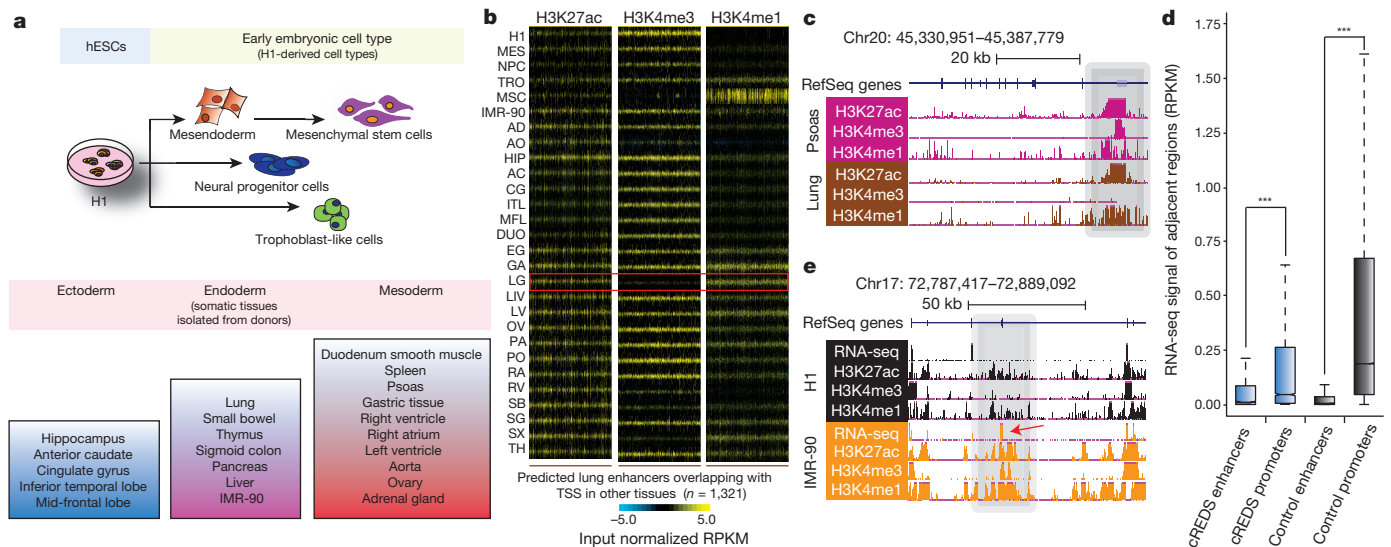


Figure 1 | Epigenome profiles of tissues reveal cREDS with dynamic histone modification signatures. **a**, Schematic of the cell/tissue types profiled and their progression along developmental lineages. Samples include embryonic stem cells (H1), early embryonic lineages (mesendoderm cells (MES), neural progenitor cells (NPC), trophoblast-like cells (TRO) and mesenchymal stem cells (MSC)) and somatic primary tissues, representative of all three germ layers (ectoderm: hippocampus (HIP), anterior caudate (AC), cingulate gyrus (CG), inferior temporal lobe (ITL) and mid-frontal lobe (MFL); endoderm: lung (LG), small bowel (SB), thymus (TH), sigmoid colon (SG), pancreas (PA), liver (LIV) and IMR-90 fibroblasts; mesoderm: duodenum smooth muscle (DUO), spleen (SX), psoas (PO), gastric tissue (GA), right heart ventricle (RV), right heart atrium (RA), left heart ventricle (LV), aorta (AO), ovary (OV) and

adrenal gland (AD)). **b**, Heat maps show H3K27ac, H3K4me3 and H3K4me1 enrichment (input normalized reads per kilobase per million mapped reads (RPKM)) at predicted lung enhancers ($n = 1,321$), which are defined as promoters in other tissues, across all 28 samples. The red box highlights the signatures in lung. **c**, A UCSC genome browser snapshot of a region on chromosome 20, showing the chromatin states of a cREDS element (grey shading) predicted as a promoter in psoas and an enhancer in lung. **d**, A box plot of RNA-seq signals (RPKM) overlapping ± 1 kb of cREDS enhancers, cREDS promoters, non-cREDS control enhancers and non-cREDS control promoters. *** $P < 10 \times 10^{-142}$, Wilcoxon test. **e**, RNA-seq and chromatin states of a cREDS element (grey shading) is shown for a region on chromosome 17 in H1 and IMR-90. Arrow indicates an alternative exon incorporated in IMR-90.

analysis of gene expression (CAGE) signal²¹, typical of active promoters (Extended Data Fig. 6a, b), and found that they possess heart-restricted enhancer activities in an *in vivo* zebrafish reporter assay (Extended Data Fig. 6c). Consistent with reporter activities, transcriptional properties (RNA-seq values based on reads per kilobase per million mapped reads (RPKM) within ± 1 kb of the elements) of cREDS enhancers and promoters are similar to non-cREDS enhancers and promoters, respectively (Fig. 1d). Interestingly, when comparing isoform dynamics across H1 and IMR-90 RNA-seq data sets²² with cREDS identified between these two cell types we discovered that a subset of cREDS promoters was accompanied by creation of new transcripts and/or alternative exon usage ($n = 99$) (Fig. 1e), revealing a possible function whereby cREDS influence cell/tissue-specific transcript variants. Taken together, these data show that cREDS can potentially function as both promoters and enhancers in distinct cell types and fine-tune transcriptomes.

Reasoning that global analysis of allelic histone modification and gene expression patterns would elucidate mechanisms of long-range gene regulation by distal *cis*-regulatory elements, we re-analysed RNA-seq and ChIP-seq data sets by considering haplotype information. For this purpose, we applied HaploSeq¹, which integrated genome sequencing with high-throughput chromatin conformation capture (Hi-C) data sets to derive chromosome-spanning haplotypes (see Supplementary Information). For four different tissue donors, we generated haplotypes spanning entire chromosomes with 99.5% completeness on average (the coverage of haplotype-resolved genomic regions) and average resolution (the coverage of phased heterozygous SNPs) ranging from 78% to 89% (Fig. 2a and Supplementary Tables 4 and 5). The accuracy of haplotype predictions was validated by the concordance with SNPs residing in the same paired-end sequencing reads. The concordance rates were 99.7% and 98.4% for H3K27ac ChIP-seq reads (described below) and RNA-seq reads, respectively, indicating high accuracy. We then re-analysed 36 mRNA-seq data sets from 18 tissues (including 16 tissues noted above with the addition of bladder and adipose tissue) and 187 ChIP-seq data

sets for 6 histone modifications (Supplementary Table 6), from up to 4 individual donors, in a haplotype-resolved context.

Although widespread allelic imbalances in gene expression had been previously noted^{17,23–25}, it remains unclear whether this phenomenon is consistent across distinct tissues and individuals, and the underlying mechanism remains undefined. To address the first point, we defined genes with allelically biased expression by means of mapping the RNA-seq reads in each tissue sample in a haplotype-resolved manner. We observed extensive allelically biased gene expression, ranging from 4% to 13% of all informative genes (>10 allelic read counts) in each tissue sample (false discovery rate (FDR) = 5%, Extended Data Fig. 7a, b). Comparatively, the proportion of allelically biased genes in individual tissue donors ranged from 6% to 23% of all informative genes, giving a combined total of 2,570 allelically biased genes (Fig. 2b and Supplementary Table 7). As a control, known imprinted genes ($n = 15$) showed common allelic biases across multiple samples (Fig. 2c) and donors (Extended Data Fig. 7c). Our data sets, representing the only collection of haplotype-resolved transcriptomes across an array of tissues from multiple individuals, allowed us to characterize allelic transcription across tissues and donors. While most genes with allelically biased expression demonstrate bias in multiple samples, approximately 75% exhibit statistically significant donor-specific bias (Fig. 2d and Extended Data Fig. 7d). This suggests a connection between sequence differences of individuals and allelically biased gene expression. In support of this model, genes frequently demonstrate consistent direction of allelic bias across multiple tissues of a given donor (Fig. 2e and Extended Data Fig. 7e). Interestingly, allelically biased genes were not restricted to the same tissue type across distinct donors. Rather, they were mostly specific to individual samples derived from each donor (Fig. 2f and Extended Data Fig. 7f), possibly resulting from differential levels of tissue-restricted transcription factors among different tissue samples.

As natural genetic variations can affect enhancer selection and function in mammalian cells²⁶, we hypothesized that polymorphisms at

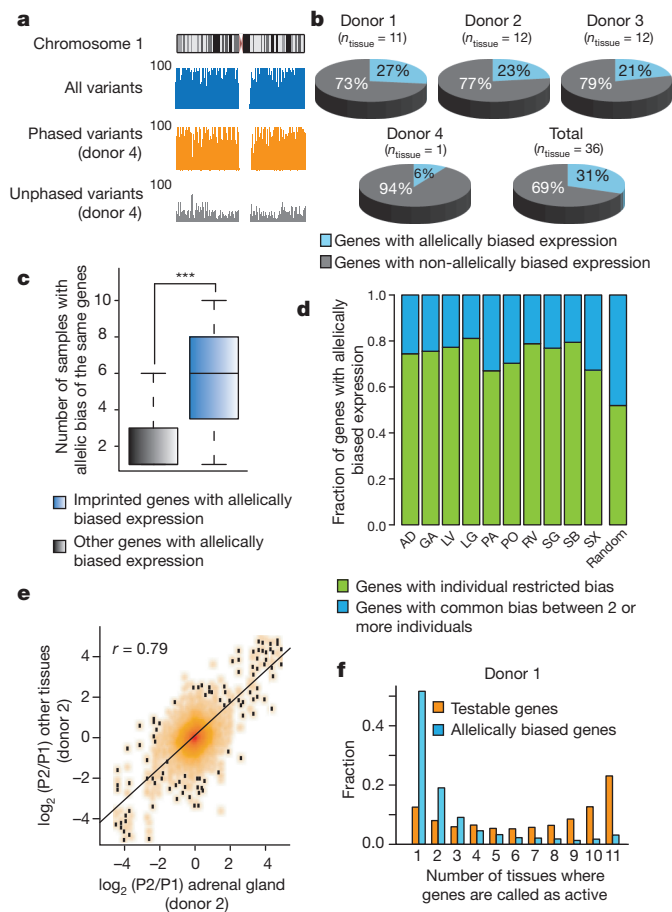


Figure 2 | Widespread, individual-specific allelic bias in gene expression.

a, Genome browser snapshots illustrate completeness and resolution of haplotypes resolved in donor 4. The y axis indicates the number of variants within 100-kb windows. The density of all (blue), phased (orange) and unphased (grey) variants across chromosome 1 is shown. **b**, Proportion of genes with allelically biased expression among informative genes and the number of tissue samples derived from each donor (n_{tissue}) are described. **c**, Box-plot illustrates occurrence of imprinted ($n = 15$) and other allelically biased genes, excluding pseudogenes ($n = 2,334$) across samples. $***P < 9.9 \times 10^{-5}$, Kolmogorov–Smirnov (KS) test. **d**, Including only tissues with two or three equivalent samples derived from distinct donors ($n_{\text{tissue}} = 10$), genes with allelic imbalances were defined as common between individuals (consistent bias among same tissue type from multiple donors) or as individual-restricted. Random control represents average from randomly selected samples (10,000 iterations). Abbreviations are defined in the legend to Fig. 1. **e**, Fold change of gene expressions between alleles (parental allele 1 (P1) and 2 (P2)) in adrenal gland from donor 2 (x axis) is compared to all other tissues from donor 2 (y axis). **f**, A histogram illustrates the proportions of allelically expressed genes in donor 1 ($n = 1,375$) defined in various numbers of tissues. The fraction of all testable genes or allelically expressed genes (y axis) is calculated for the number of tissues where they are identified as active (x axis) (P value $< 2.2 \times 10^{-16}$, KS test).

(H3K36me3) (see Supplementary Information). In support of our hypothesis, the allelic biases of gene expression strongly agreed with chromatin states of sequences at or near the genes (Fig. 3a, b and Extended Data Fig. 8a).

Furthermore, if allelic imbalances of enhancer activities indeed contribute to allelically biased gene expression, we expect that chromatin states at enhancers will be concordant with the expression of their targets. Therefore, we generated additional H3K27ac ChIP-seq data sets with deeper coverage and longer sequencing reads (for better delineation of alleles) for 14 of the previously analysed tissue samples and an additional 6 samples from independent donors (Supplementary Table 7). Of the informative enhancers (with > 10 polymorphism-bearing sequence reads), 11.6% ($n = 11,714$, FDR = 1%) showed significant allelically biased H3K27ac enrichment in any tissue types (Fig. 3c and Supplementary Table 8). H3K27ac biases were validated by allele-specific ChIP-qPCR (Extended Data Fig. 8b). Interestingly, identical genotypes often yielded the same direction of biases in allelic enhancer activities (Fig. 3d). We further tested whether sequence variations are systematically associated

cis-regulatory sequences underlie the widespread allelic transcriptional biases. We thus exploited the unique resource of 187 haplotype-resolved ChIP-seq data sets to analyse the state of *cis*-regulatory elements. We identified allelically biased marks at promoter regions (H3K27ac, H3K4me1, H3K4me3, H3K27me3 and H3K9me3) and transcribed gene bodies

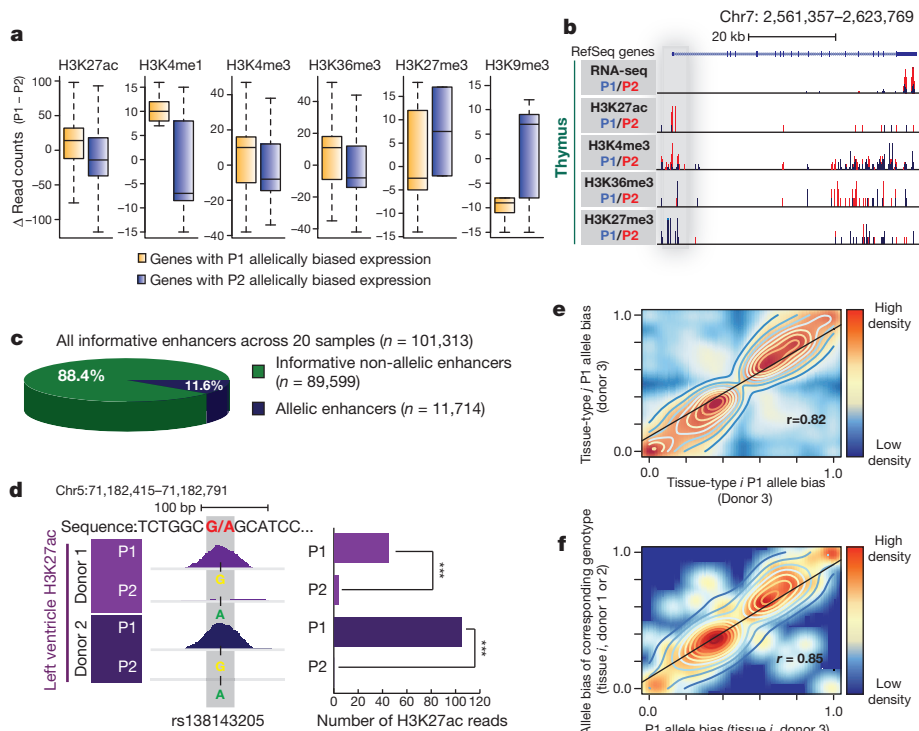


Figure 3 | Characterization of allele bias in chromatin states at *cis*-regulatory elements.

a, Box-plots present haplotype-resolved ChIP-seq reads at promoter or gene bodies (H3K27ac: $n = 744$, $P = 10 \times 10^{-14}$; H3K4me1: $n = 32$, $P = 0.035$; H3K4me3: $n = 177$, $P = 0.0047$; H3K27me3: $n = 12$, $P = 0.43$; H3K9me3: $n = 27$, $P = 0.13$; H3K36me3: $n = 291$, $P = 4.3 \times 10^{-6}$, KS test). **b**, Allelically biased gene expression of *IQCE* is concordant with chromatin marks at the promoter (grey) and gene body. **c**, Proportion of allelic ($n = 11,714$) and non-allelic ($n = 89,599$) among all informative enhancers ($n = 101,313$) across 20 tissue samples. **d**, A snapshot showing a SNP (rs138143205) with H3K27ac bias towards the G allele in both left heart ventricle donors (left). Bar chart illustrates the number of H3K27ac reads corresponding to the P1 versus P2 alleles in both donors (right). $***P < 10 \times 10^{-19}$, binomial test. **e**, **f**, Scatter plots show strong correlation of the P1 allele bias of enhancer activities among two different tissue types from donor 3 ($n = 4,427$) (**e**) and among the P1 allele bias in donor 3 (x axis) and the allele bias of corresponding genotypes in donor 1 or 2 (y axis) at allelic enhancer in the same tissue type (**f**) ($n = 447$).

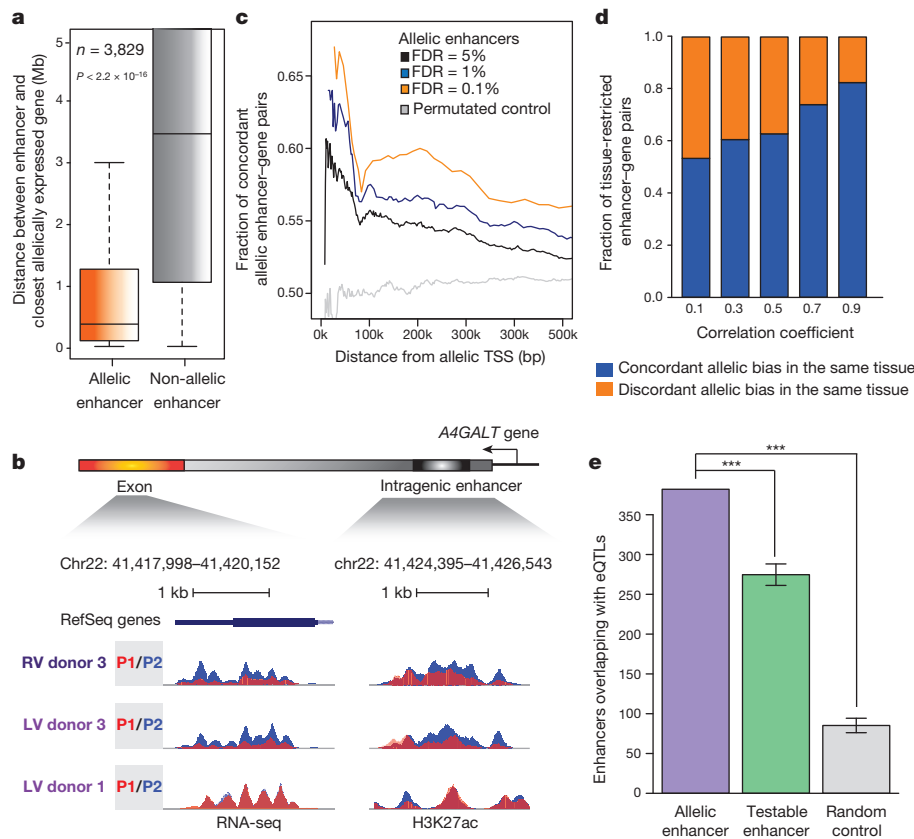


Figure 4 | Allelic histone acetylation at enhancers is associated with allelically biased gene expression. **a**, Average distance of allelic (5% FDR) and non-allelic enhancers to the closest allelically expressed gene is significantly different ($n = 3,829$, $P < 2.2 \times 10^{-16}$, KS test). **b**, Genome browser snapshots show an allelic enhancer within the intron of the allelically expressed *A4GALT* gene (P1, red; P2, blue) on chromosome 22 across three samples. **c**, Line plot presents the fraction of concordant allelic bias between allelically expressed genes and allelic enhancers in terms of distance. The allelic enhancer-gene pairs were defined with FDR cutoff values of 5% ($n = 14,082$) (black), 1% ($n = 6,057$) (blue) and 0.1% ($n = 2,362$) (orange). Permutated control of a set of enhancer-gene pairs was included ($n = 14,082$) (grey). Distance between allelically biased enhancer-gene pairs and fraction of concordant allelic bias are denoted by x and y axes, respectively ($P < 2.2 \times 10^{-16}$, KS test). **d**, Fractions of tissue-restricted enhancer-gene pairs ($n = 3,106$) (y axis) that show concordant (blue) or discordant (orange) allelic biases in the same tissue, are presented across a range of Pearson correlation coefficients (x axis) ($P < 2.2 \times 10^{-16}$, KS test, random permutated control concordant pairs = 50%). **e**, Overlap between eQTLs³⁰ and allelic enhancers; testable enhancers or random control regions are shown. Error bars represent standard deviations. Testable enhancers and random control regions were generated 10,000 times with the same numbers as allelic enhancers. *** $P < 10 \times 10^{-5}$.

with allelic H3K27ac, which reflects enhancer activities²⁷. Indeed, H3K27ac biases were strongly correlated with specific genotypes, whereby given identical genotypes, this histone modification was biased to the same alleles, both across tissue types and individuals (Fig. 3d–f and Extended Data Fig. 9a). Furthering this finding, we analysed previously generated data sets from lymphoblastoid cell lines²⁸ and found similar significant correlation of genotype and molecular phenotype of H3K27ac

enrichment (Extended Data Fig. 9b). Taken together, these data reveal that extensive allelic imbalance events are associated with sequence variants in *cis*-regulatory elements.

We discovered that allelic enhancers resided in significantly closer proximity to genes with allelically biased expression, as compared to non-allelic enhancers (Fig. 4a, b). We also observed examples where distinct tissues from the same donor showed similar allelic biases of gene

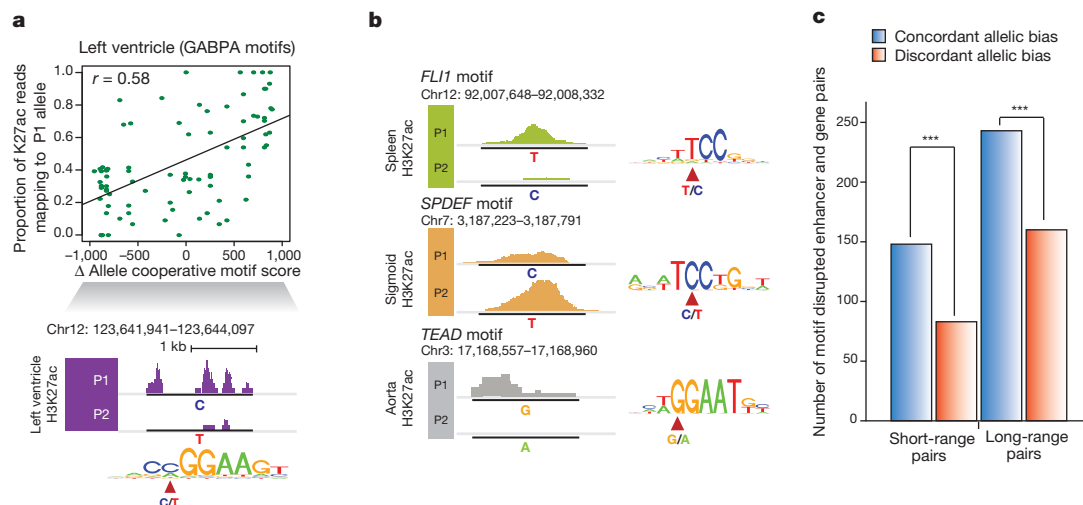


Figure 5 | Motif disruption by genetic variants is concordant with allelic H3K27ac biases at enhancers. **a**, Differential GABPA binding motif scores between two alleles (P1 – P2 motif scores) in left heart ventricle are correlated with the proportion of H3K27ac reads corresponding to the P1 allele (top). Values range from negative to positive, indicating P1 and P2 motif disruption, respectively. An example on chromosome 12 illustrating that P1, with a motif preserving C allele, has higher H3K27ac enrichment and that P2, with the motif disrupting T allele, has little H3K27ac enrichment (bottom). **b**, Three examples (*FLI1* in spleen, *SPDEF* in sigmoid colon, and *TEAD* in aorta) of

motif-disrupted enhancers demonstrate allelic biased activities. The variant location and genotypes of P1 and P2 alleles are marked in motif sequence. **c**, All possible motif disrupted enhancer-gene pairs within the indicated distance window are defined with concordant allelic bias (blue, gene-enhancer pairs with biases towards the same allele) or discordant allelic bias (red, gene-enhancer pairs with biases towards different alleles). Only thymus, left heart ventricle and aorta were considered due to the availability of Hi-C data. Short-range pairs are defined if any allelically expressed genes are located <20 kb away. *** $P < 2.5 \times 10^{-5}$, binomial test.

expression and H3K27ac at enhancers (left ventricle and right ventricle from donor 3); however, the same tissue type derived from a different donor (left ventricle from donor 1) yielded no consistent patterns (Fig. 4b), supporting the hypothesis that allelically biased gene expression is driven by individual-specific genetic variation in enhancers. Indeed, within close proximity, the concordance between allelic enhancers and gene expression is significantly higher than permuted control enhancer/gene sets (Fig. 4c). Remarkably, 56% of allelic enhancer–gene pairs are greater than 300 kb apart (Extended Data Fig. 10a, b), the delineation of which was enabled by whole-chromosome-spanning haplotypes.

Similar to genes, many allelically biased enhancers are tissue restricted (Extended Data Fig. 10c). We reasoned that gene expression biases could result from tissue-restricted enhancer activities, evidenced by significant correlation between allelic enhancers and allelically expressed genes (Fig. 4d). Allelic enhancers also significantly overlapped with expression quantitative trait loci (eQTLs) (Fig. 4e), DNase I hypersensitivity QTLs and H3K27ac QTLs (Extended Data Fig. 10d), defined independently^{28–30}, corroborating the functional roles of identified allelic enhancers on gene regulation. Taken together, these observations support a model whereby allelic biases of *cis*-regulatory element activities could be responsible for allelic gene expression.

Finally, to elucidate further the mechanism by which allelically biased enhancer activities arise, we examined single nucleotide polymorphisms (SNPs) that potentially disrupt or weaken transcription factor binding motifs. We calculated changes in motif score between alleles (motif disruption score) at allelic enhancers and discovered 133 transcription factor motifs showing significant concordance between allelic reduction of enhancer activities and transcription factor motif disruption (Fig. 5a, b) (FDR = 10%, Supplementary Table 9). Moreover, genes with allelically biased expression were concordant with enhancer motif disruptions within close proximity (<20 kb) or displaying strong Hi-C interactions at longer distances (>20 kb) (Fig. 5c and Supplementary Information). Our results therefore suggest that genetic variations are probably responsible for allelic enhancer activities and consequently allelically biased gene expression.

By generating chromosome-spanning haplotypes, we carried out a comprehensive survey of allelic chromatin state and gene expression. We found evidence for extensive allelically biased gene expression, which is connected to change in chromatin states at *cis*-regulatory elements, probably resulting from transcription factor binding disruption by sequence variations. These observations mirror findings in mice where allelic biases of *cis*-regulatory element activities could be responsible for allelic gene expression²⁶, and demonstrate that such a phenomenon is probably widespread in the human genome. These observations shed light on the importance of considering genetic variants in understanding individual-specific gene regulation. Analyses of haplotype-resolved transcriptomes and epigenomes in additional individuals and tissues should further illuminate the role of sequence variations in defining individual-specific transcriptional programs and phenotypes.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 25 November 2013; accepted 7 January 2015.

- Selvaraj, S., Dixon, J. R., Bansal, V. & Ren, B. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nature Biotechnol.* **31**, 1111–1118 (2013).
- Xie, W. *et al.* Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell* **153**, 1134–1148 (2013).
- Zhu, J. *et al.* Genome-wide chromatin state transitions associated with developmental and environmental cues. *Cell* **152**, 642–654 (2013).
- Rivera, C. M. & Ren, B. Mapping human epigenomes. *Cell* **155**, 39–55 (2013).
- Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genet.* **39**, 311–318 (2007).
- Rajagopal, N. *et al.* RFECS: a random-forest based algorithm for enhancer identification from chromatin state. *PLOS Comput. Biol.* **9**, e1002968 (2013).
- Fang, F. *et al.* Genomic landscape of human allele-specific DNA methylation. *Proc. Natl Acad. Sci. USA* **109**, 7332–7337 (2012).

- Stergachis, A. B. *et al.* Developmental fate and cellular maturity encoded in human regulatory DNA landscapes. *Cell* **154**, 888–903 (2013).
- Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
- Flandez, M. *et al.* Nr5a2 heterozygosity sensitizes to, and cooperates with, inflammation in KRas(G12V)-driven pancreatic tumorigenesis. *Gut* **63**, 647–655 (2014).
- Hirai, H. *et al.* Involvement of Runx1 in the down-regulation of fetal liver kinase-1 expression during transition of endothelial cells to hematopoietic cells. *Blood* **106**, 1948–1955 (2005).
- Hwang, D. H. *et al.* Transplantation of human neural stem cells transduced with Olig2 transcription factor improves locomotor recovery and enhances myelination in the white matter of rat spinal cord following contusive injury. *BMC Neurosci.* **10**, 117 (2009).
- Jahan, I., Kersigo, J., Pan, N. & Fritzsche, B. Neurod1 regulates survival and formation of connections in mouse ear and brain. *Cell Tissue Res.* **341**, 95–110 (2010).
- Lee, C. S. *et al.* Loss of nuclear factor E2-related factor 1 in the brain leads to dysregulation of proteasome gene expression and neurodegeneration. *Proc. Natl Acad. Sci. USA* **108**, 8408–8413 (2011).
- Moya, M. *et al.* Foxa1 reduces lipid accumulation in human hepatocytes and is down-regulated in nonalcoholic fatty liver. *PLoS ONE* **7**, e30014 (2012).
- Kunarski, G. *et al.* Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nature Genet.* **42**, 631–634 (2010).
- Xie, M. *et al.* DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape. *Nature Genet.* **45**, 836–841 (2013).
- Lu, X. *et al.* The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. *Nature Struct. Mol. Biol.* **21**, 423–425 (2014).
- Kowalczyk, M. S. *et al.* Intragenic enhancers act as alternative promoters. *Mol. Cell* **45**, 447–458 (2012).
- Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser—a database of tissue-specific human enhancers. *Nucleic Acids Res.* **35**, D88–D92 (2007).
- The FANTOM Consortium and the RIKEN PMI and CLST (DGT). A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014).
- Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotechnol.* **31**, 46–53 (2013).
- Gimelbrant, A., Hutchinson, J. N., Thompson, B. R. & Chess, A. Widespread monoallelic expression on human autosomes. *Science* **318**, 1136–1140 (2007).
- Kilpinen, H. *et al.* Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science* **342**, 744–747 (2013).
- Skelly, D. A., Johansson, M., Madeoy, J., Wakefield, J. & Akey, J. M. A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. *Genome Res.* **21**, 1728–1737 (2011).
- Heinz, S. *et al.* Effect of natural genetic variation on enhancer selection and function. *Nature* **503**, 487–492 (2013).
- Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl Acad. Sci. USA* **107**, 21931–21936 (2010).
- McVicker, G. *et al.* Identification of genetic variants that affect histone modifications in human cells. *Science* **342**, 747–749 (2013).
- Kasowski, M. *et al.* Extensive variation in chromatin states across humans. *Science* **342**, 750–752 (2013).
- Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).

Supplementary Information is available in the online version of the paper.

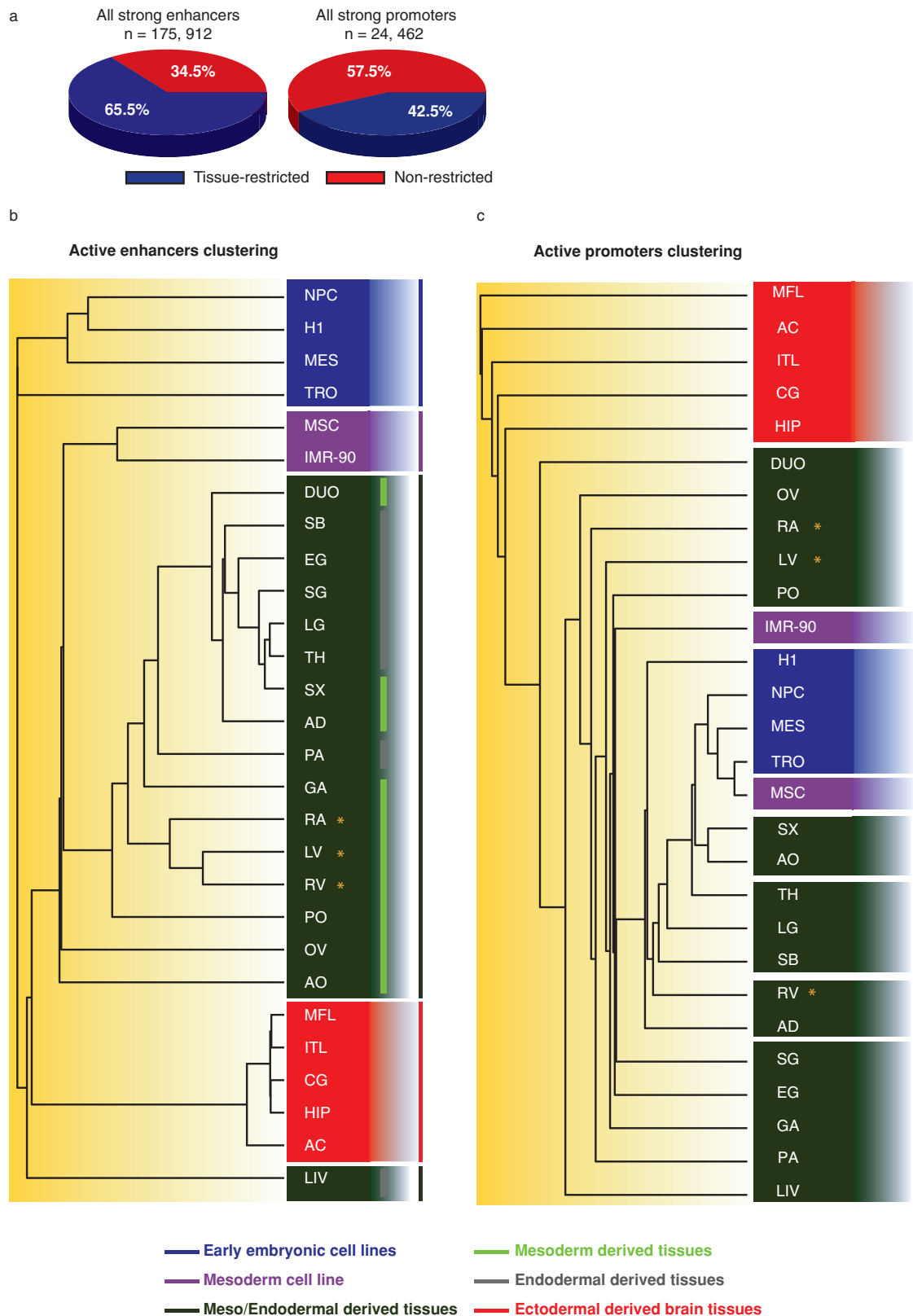
Acknowledgements This work is supported by the NIH Epigenome Roadmap Project (U01 ES017166), CIRM RN2-00905-1, NIH ES017166, NSFC 91019016, NBRPC 2012CB316503 and NIH Fellowship Grants F32HL110473 and K99HL119617. We thank A. Kulkarni and J. Wu for help with processing RNA-seq data sets, and Y. He and M. Schultz for discussions regarding allelic analyses of RNA-seq data sets. We also thank members of the Ren laboratory for comments.

Author Contributions D.L., W.X., J.R.E., N.C.C. and B.R. led the data production. I.J., N.R., M.Q.Z., J.R.E. and B.R. led the data analyses. I.J., N.R., S.S., F.Y., Y.Q., L.E., M.H., A.S. and P.R. conducted analyses. S.L. and Y.L. processed tissue samples. D.L., A.S., A.Y.L., C.-A.Y., S.K. and H.Y. produced data. D.L., I.J., N.R. and B.R. wrote the manuscript.

Author Information ChIP-seq and RNA-seq data sets were deposited at the Gene Expression Omnibus (GEO) under accession number GSE16256. Hi-C data sets were deposited at GEO under accession number GSE58752. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to B.R. (biren@ucsd.edu).



This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported licence. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons licence, users will need to obtain permission from the licence holder to reproduce the material. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-sa/3.0>

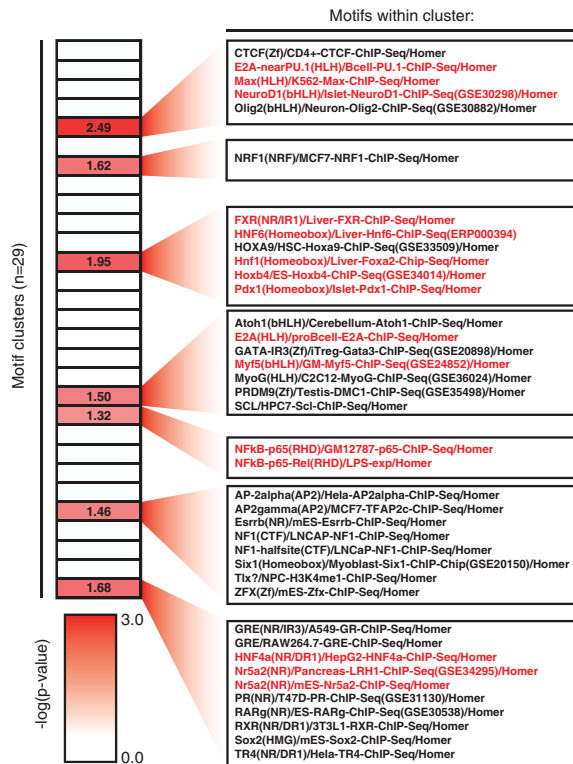


Extended Data Figure 1 | Active enhancers cluster along developmental lineages. **a**, Pie charts showing fractions of tissue-restricted and non-tissue-restricted strong enhancers and promoters. **b**, Hierarchical clustering with optimal leaf ordering based on all H3K27ac-marked highly active enhancers. Four major clusters are represented: early embryonic cell types (blue); a large set of meso/endoderm-derived tissues (dark green); a set consisting of ectoderm-derived brain tissues (red); and a small cluster of mesoderm cell lines (purple), which bridged the early embryonic lineages with the somatic tissues.

Although TRO did not fall within any clusters, it shared the highest degree of similarity with the early embryonic cell lines. On a subsequent level, two clusters are seen separating endoderm-derived tissues (grey line) and mesoderm-derived tissues (green line). Heart tissues are denoted by yellow asterisks. **c**, Clustering of tissues by promoters' histone acetylation status shows grouping of tissues that are of similar types but are less evident in germ-layer divisions than clustering of enhancers.

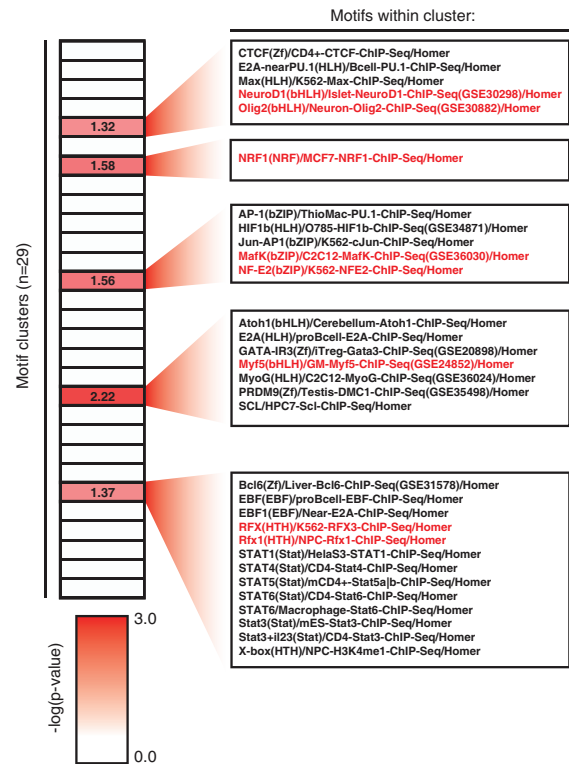
a

Enrichment of Pancreas-restricted enhancers



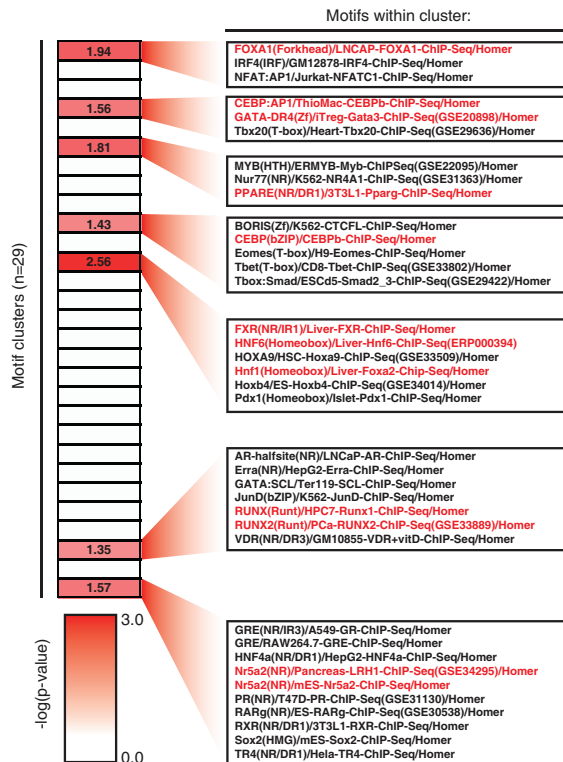
b

Enrichment of Anterior Caudate-restricted enhancers



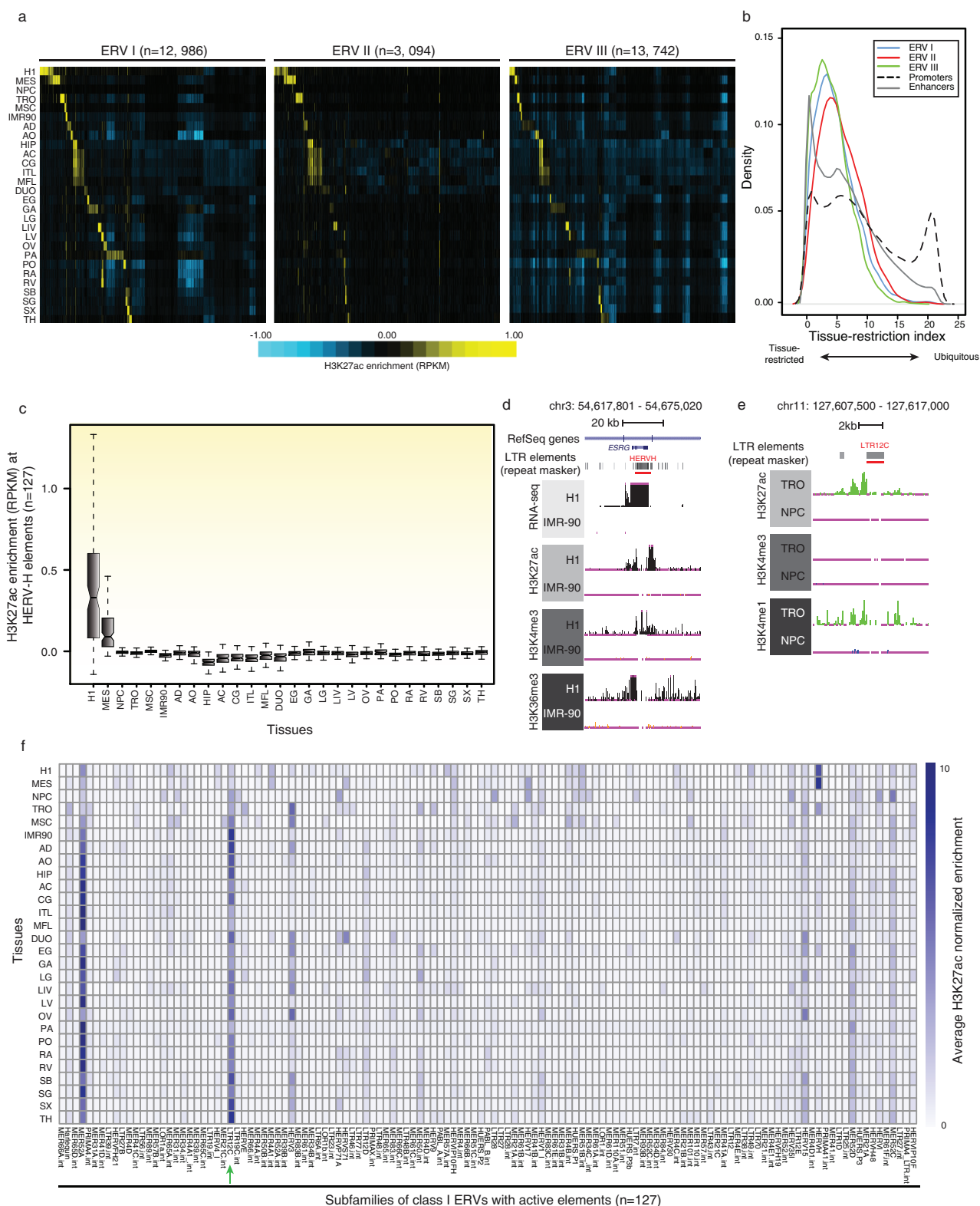
c

Enrichment of Liver-restricted enhancers



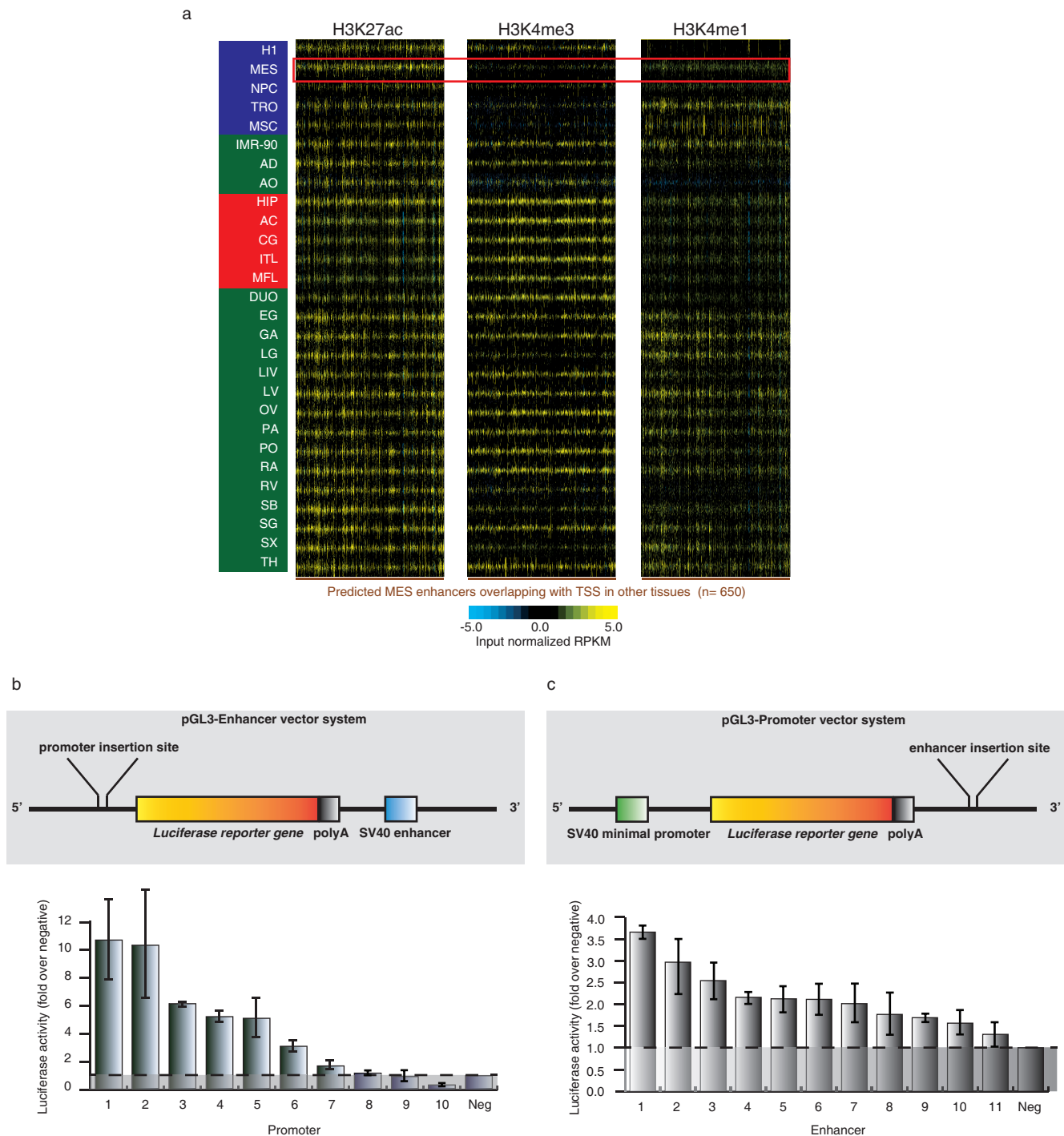
Extended Data Figure 2 | Tissue-restricted enhancers are enriched for transcription factor motifs important for cell identity and/or function. Significantly enriched motifs ($P < 10 \times 10^{-10}$) across all 28 tissues are divided into 29 clusters (method described in Supplementary Information). An overall P value is generated for the enrichment of each tissue for each cluster. The figure illustrates $-\log(P$ value) of (a) pancreas, (b) anterior caudate and (c)

liver-restricted enhancer motif enrichment for the various clusters. For ease of visualization, any cluster with P values greater than 0.05 is denoted 0. Red highlighted text refers to a subset of motifs for transcription factors with literature support (see Supplementary Information) to have function in (a) the pancreas, (b) the brain and (c) the liver.



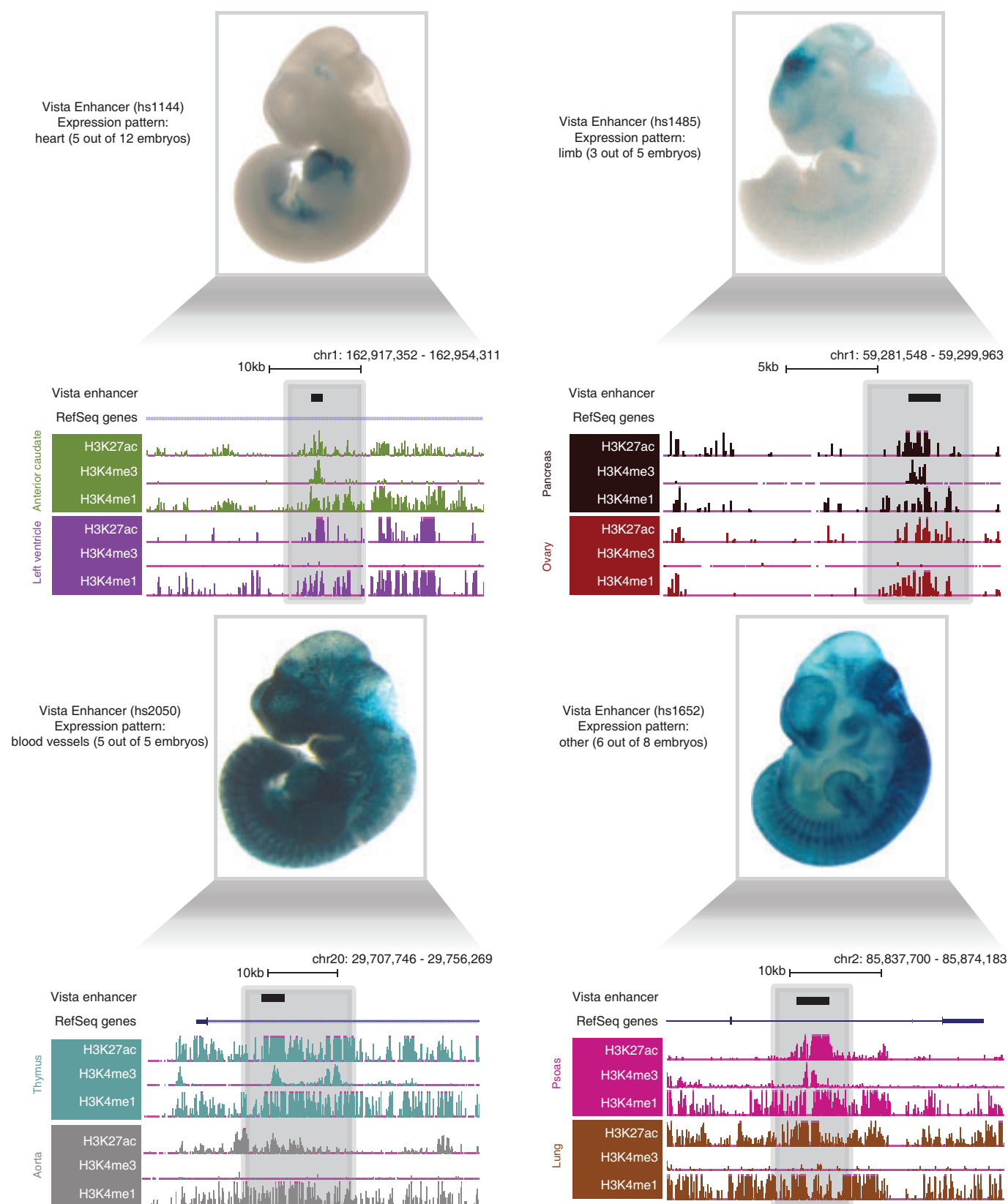
Extended Data Figure 3 | Endogenous retroviruses (ERVs) are enriched for active *cis*-regulatory element marks in a tissue-restricted fashion. **a**, A clustered heat map showing the H3K27ac enrichment (RPKM) of all mappable elements of the three classes of ERVs. **b**, Distribution of the Shannon entropy of H3K27ac across enhancers, promoters and three classes of ERVs is shown as a density curve, demonstrating that H3K27ac enrichment of ERVs is more tissue-restricted than promoters and slightly less than enhancers. **c**, Box-plots illustrating the H3K27ac enrichment of 127 mappable members of the HERV-H subfamily across all tissue/cell types. The enrichment in H1 hESCs is significantly higher than all other cell/tissues types ($P < 1.4 \times 10^{-9}$, Wilcoxon

test). **d**, UCSC genome browser snapshots showing example of an HERV-H element harbouring H1-restricted active promoter marks, corresponding RNA-seq signal and H3K36me3 enrichment. Notably, this particular element has been annotated in RefSeq as the ES cell related gene (*ESRG*), a human-specific long non-coding RNA gene. **e**, UCSC genome browser snapshots showing example of a LTR12C element harbouring TRO-restricted active enhancer chromatin marks. **f**, A matrix illustrating the average H3K27ac enrichment for subfamilies of class I ERVs across all cell and tissue types. LTR12C subfamily (green arrow) shows enrichment of H3K27ac across many distinct cell types and tissues.



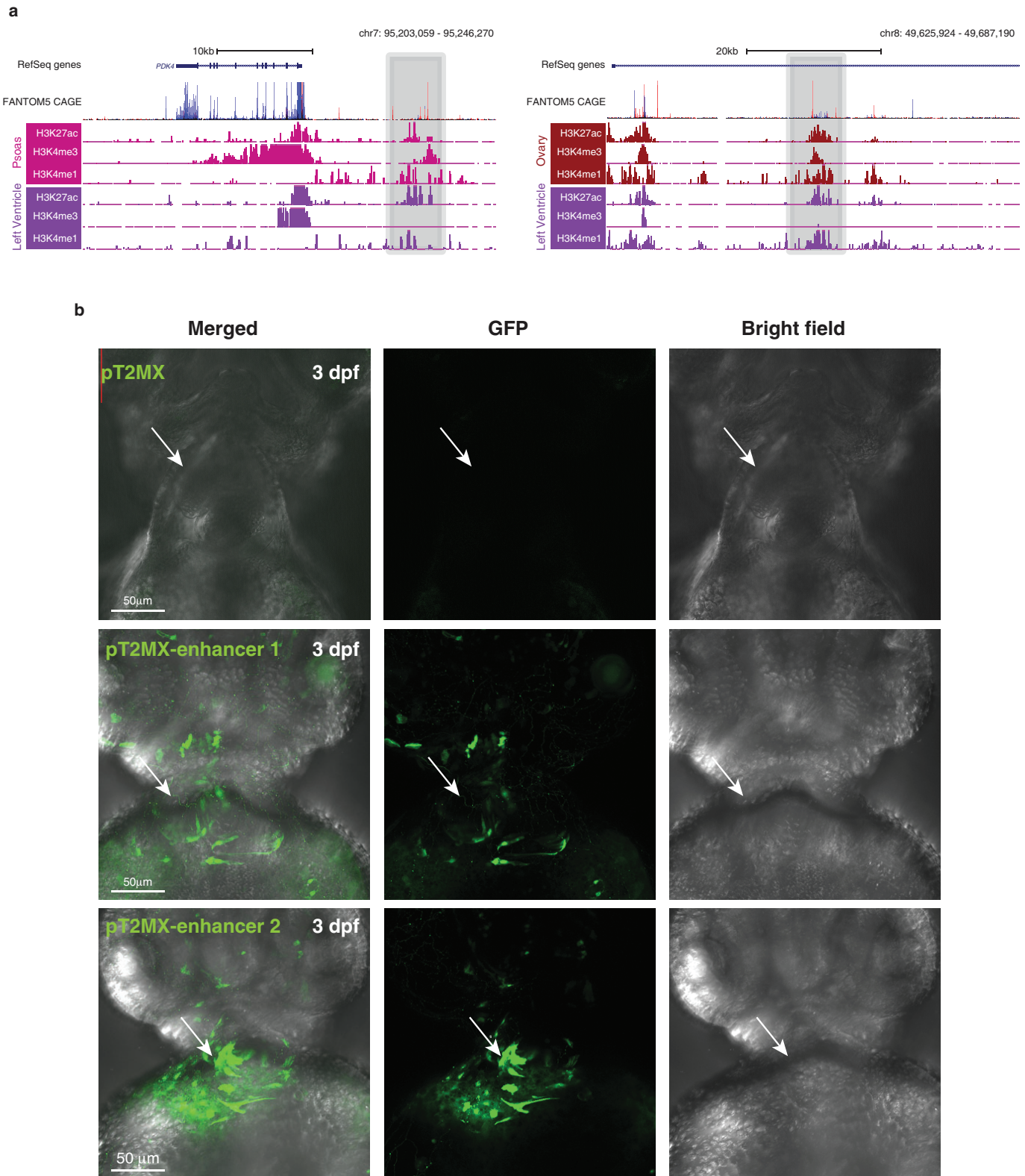
Extended Data Figure 4 | cREDS are enriched with dynamic histone mark signatures in different tissues and have putative *cis*-regulatory functions. **a**, Heat maps showing the enrichment (RPKM) of the H3K27ac, H3K4me3 and H3K4me1 at MES-restricted enhancers ($n = 650$), which are predicted as promoters in other tissues, across all 28 samples. The red box highlights the histone modifications in MES cells. **b**, A schematic of the pGL3-enhancer vector used in these luciferase-reporter assays (top) and the activity of 10 selected cREDS with promoter signatures and a negative control region cloned

5' to the reporter gene after transfection into H1 hESCs (bottom). Luciferase activity of each region is normalized to the average activity of the negative controls. **c**, A schematic of the pGL3 promoter vector used in these luciferase-reporter assays (top) and the activity of 11 selected cREDS with enhancer signatures and a negative control region cloned 3' to the reporter gene after transfection into H1 hESCs (bottom). Luciferase activity of each region is normalized to averaged activity of negative control regions. Error bars reflect standard deviation between three technical replicates.



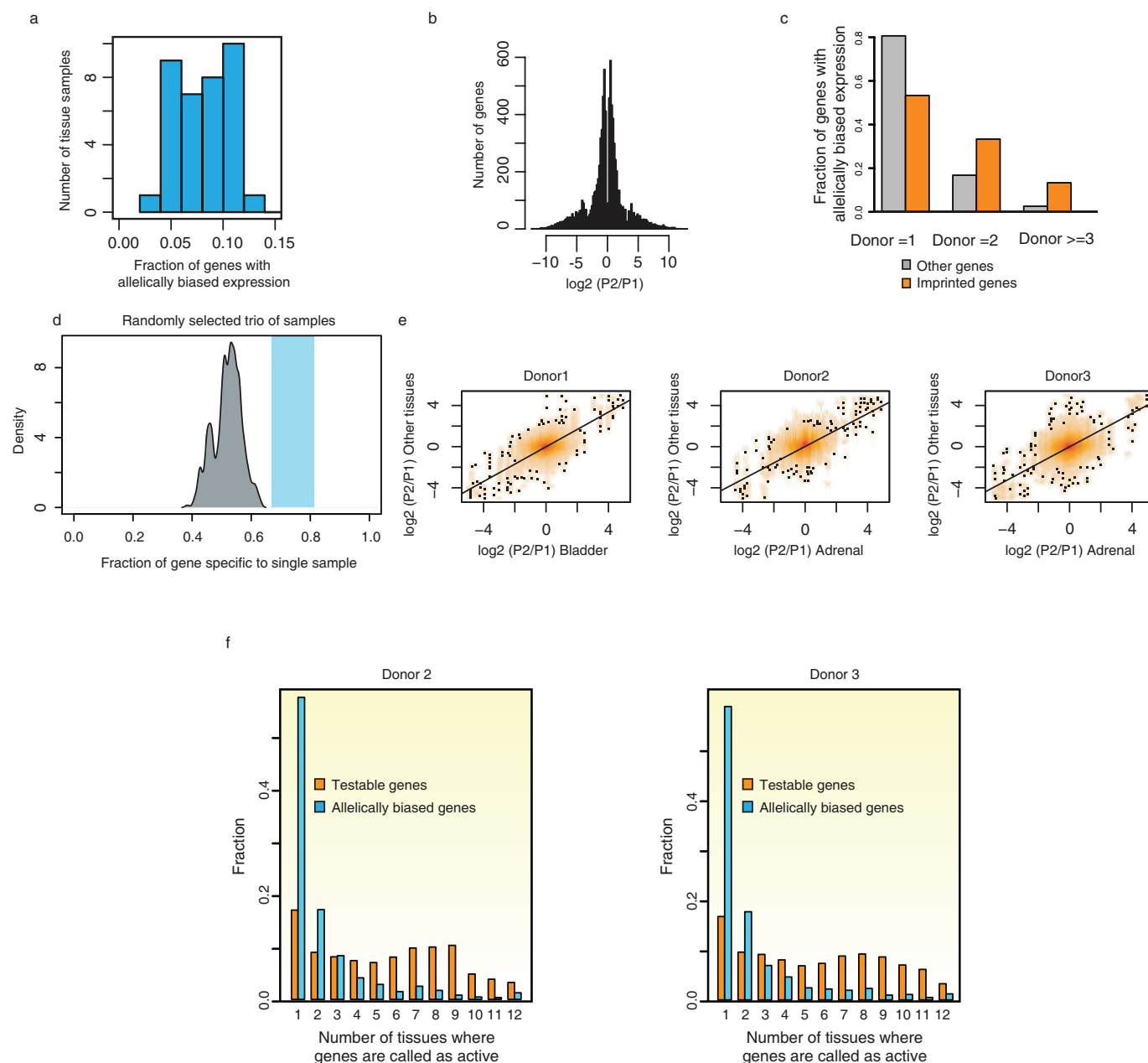
Extended Data Figure 5 | VISTA validated enhancers also possess dynamic histone modification signatures across tissues. Example screen shots of VISTA validated enhancers and the patterns of activity *in vivo* are displayed

along with histone modification patterns in representative tissues (adapted from VISTA enhancer browser²⁰).



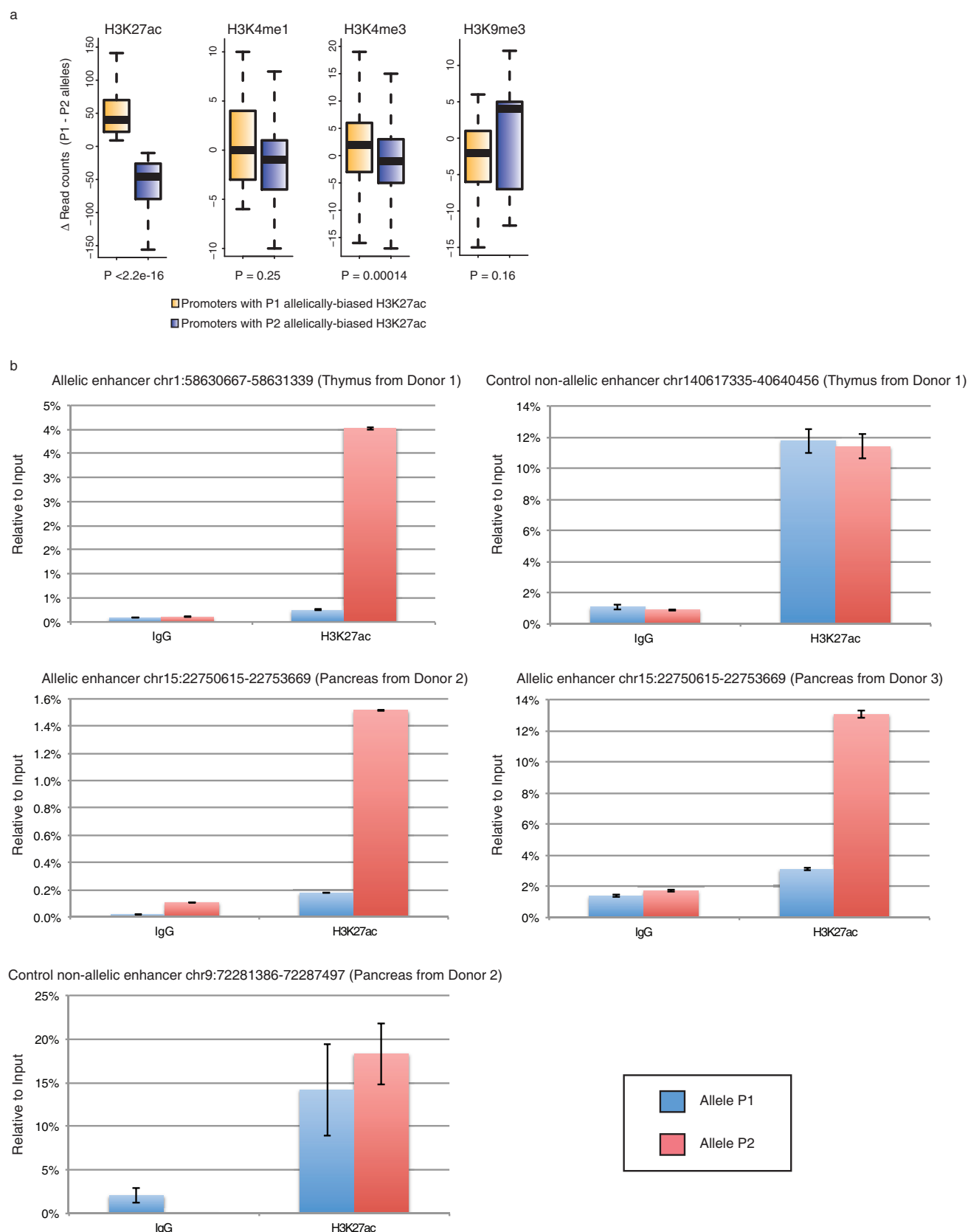
Extended Data Figure 6 | cREDS show enrichment of CAGE signal and putative enhancer functions in a zebrafish reporter assay. a, UCSC genome browser screen shots show the two cREDS elements (grey shading) harbouring enhancer and promoter signatures in distinct tissue types. When compared to CAGE data sets from the FANTOM5 project, these elements show substantial overlap with transcript signals (red and blue signals indicate CAGE signal on the forward and reverse orientation, respectively). **b,** Selected cREDS (same elements as above) with enhancer marks in left ventricle show

heart-restricted enhancer activity, as indicated by GFP expression, in 3 days post-fertilization (3 dpf) zebrafish larvae. In parallel, pT2MX negative control did not show any GFP expression. White arrow indicates location of the 3 dpf zebrafish heart. For enhancer 1, 13 out of the 38 surviving embryos showed similar patterns. For enhancer 2, 18 out of the 35 surviving embryos showed similar patterns. None of the 30 surviving embryos, injected with the control vector, showed any appreciable GFP signal in the heart. Scale bar, 50 μ m.



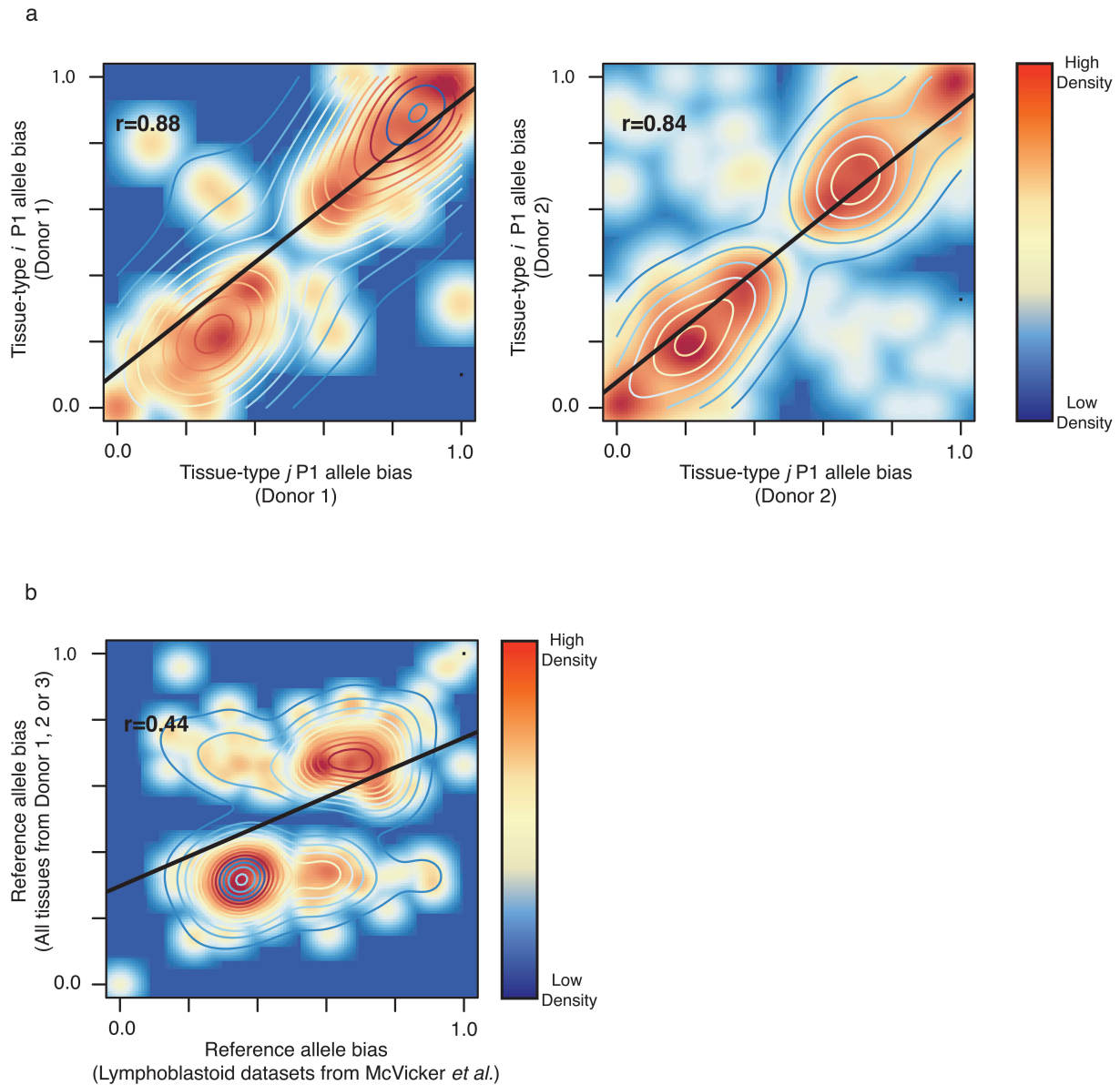
Extended Data Figure 7 | Identification of widespread allelically expressed genes. **a**, Fraction of genes with allelically biased expression in each sample. *y* axis indicates number of samples and *x* axis indicates fraction of allelically biased genes among informative genes (more than 10 SNP-containing short reads). **b**, Distribution of fold change of allelically biased genes between P1 and P2 alleles. **c**, The occurrence of allelically biased imprinted and other genes is shown. *x* axis refers to the number of individual donors where corresponding allelically expressed genes are commonly detected. **d**, A density plot showing the fraction of sample-restricted genes with allelically biased expression (grey). Three tissue samples were randomly selected and sample-restricted allelically expressed genes were defined, which includes random variance effect. The random selection was repeated 10,000 times. The shaded blue box indicates the range of fractions of individual-restricted allele-biased genes in all analysed tissue types ($n = 10$). The fraction of sample-restricted

allelically biased genes is lower than individual-restricted allele-biased genes in Fig. 2e. **e**, Fold change of allele-biased gene expression between two alleles is shown as scatter plot. *x* axis is for the fold changes in one randomly selected tissue in each donor and *y* axis is for the fold changes in all other remaining tissues in the corresponding donor. Allelic bias in one tissue is highly correlated with allelic bias in other tissues in the same individual. **f**, A histogram illustrates the proportions of allelically expressed genes in donor 2 (left) and 3 (right) defined in various numbers of tissues. The fraction of all testable genes or allelically expressed genes (*y* axis) is calculated for the number of tissues where they are called as active (*x* axis). The results indicate that the majority of allelically biased genes, as opposed to testable genes, are restricted to one or two tissue samples. KS test was performed between allele-biased genes and testable genes ($P < 2.2 \times 10^{-16}$).



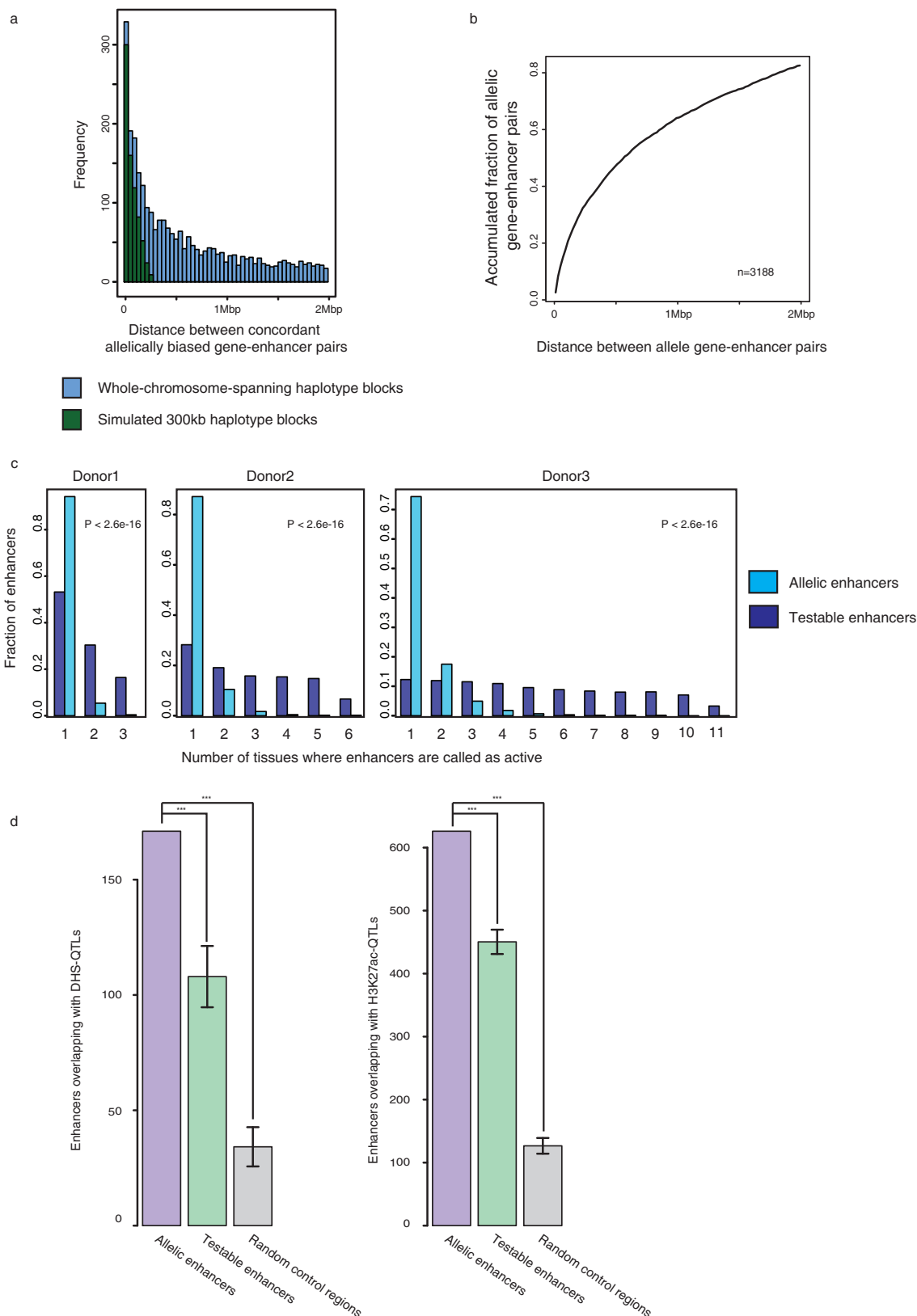
Extended Data Figure 8 | Allele-biased chromatin states. **a**, Box-plots illustrating haplotype-resolved ChIP-seq signal enrichment on the two alleles at promoter regions. The P1 or P2 allele-biased promoter regions were defined by H3K27ac signals and then H3K4m1, H3K4me3 and H3K9me3 signals were presented for the corresponding promoter regions. All chromatin states are consistent according to the allele-biased H3K27ac patterns. KS test was

performed for P value calculation. **b**, Allelically biased enhancers were tested in thymus from donor 1 and pancreas from donors 2 and 3. H3K27ac enrichment was tested by allele-specific ChIP-qPCR. Two control enhancers were included and showed to have no allelic biases in thymus or pancreas from donor 2 (top right and bottom left, respectively).



Extended Data Figure 9 | Putative enhancers with identical genotypes in different individuals exhibit similar biases in histone acetylation. **a**, Scatter plots of P1 allele-biased enhancer activities for pairwise comparison of allele-biased enhancers in donor 1 ($n = 85$) and donor 2 ($n = 4,427$). x and

y axes indicate P1 allele bias. **b**, Scatter plot of reference allele bias of enhancer activities for pairwise comparison of allele-biased enhancers in all tissues from all three donors and lymphoblastoid data sets obtained from a previous study²⁸ ($n = 309$).



Extended Data Figure 10 | Analyses of concordant allelically biased gene-enhancer pairs. **a**, Frequency of allelically expressed genes according to the distance between concordantly allele-biased enhancer-gene pairs. Blue bars represent data obtained from whole-chromosome-spanning haplotype blocks while green bars represent data obtained from simulated 300-kb haplotype blocks. 56% of enhancer-gene pairs are more than 300 kb apart. **b**, Accumulation curve showing fraction of allelically biased genes that have at least one concordantly allelic enhancer within a given distance (x axis). Up to

83% of allelically expressed genes are within 2 Mb of a concordantly biased allelic enhancer. **c**, The frequency of allele-biased enhancers in donors 1, 2 and 3. y axis indicates fraction of enhancers and x axis indicates frequency of allelically biased enhancers. KS test was performed between allele-biased enhancers and testable enhancers. **d**, Bar plots presenting the number of enhancers overlapping with DHS QTLs and H3K27ac QTLs for allelic enhancers, testable enhancers, and random control regions. *** $P < 10 \times 10^{-5}$.

Dissecting neural differentiation regulatory networks through epigenetic footprinting

Michael J. Ziller^{1,2,3*}, Reuven Edri^{4*}, Yaakey Yaffe⁴, Julie Donaghey^{1,2,3}, Ramona Pop^{1,2,3}, William Mallard^{1,3}, Robbyn Issner¹, Casey A. Gifford^{1,2,3}, Alon Goren^{1,5,6}, Jeffrey Xing¹, Hongcang Gu¹, Davide Cacchiarelli¹, Alexander M. Tsankov^{1,2,3}, Charles Epstein¹, John L. Rinn^{1,2,3}, Tarjei S. Mikkelsen¹, Oliver Kohlbacher⁷, Andreas Gnirke¹, Bradley E. Bernstein^{1,5,6}, Yechiel Elkabetz⁴§ & Alexander Meissner^{1,2,3}§

Models derived from human pluripotent stem cells that accurately recapitulate neural development *in vitro* and allow for the generation of specific neuronal subtypes are of major interest to the stem cell and biomedical community. Notch signalling, particularly through the Notch effector HES5, is a major pathway critical for the onset and maintenance of neural progenitor cells in the embryonic and adult nervous system^{1–3}. Here we report the transcriptional and epigenomic analysis of six consecutive neural progenitor cell stages derived from a *HES5::eGFP* reporter human embryonic stem cell line⁴. Using this system, we aimed to model cell-fate decisions including specification, expansion and patterning during the ontogeny of cortical neural stem and progenitor cells. In order to dissect regulatory mechanisms that orchestrate the stage-specific differentiation process, we developed a computational framework to infer key regulators of each cell-state transition based on the progressive remodelling of the epigenetic landscape and then validated these through a pooled short hairpin RNA screen. We were also able to refine our previous observations on epigenetic priming at transcription factor binding sites and suggest here that they are mediated by combinations of core and stage-specific factors. Taken together, we demonstrate the utility of our system and outline a general framework, not limited to the context of the neural lineage, to dissect regulatory circuits of differentiation.

We used the human embryonic stem (ES) cell line WA9 (also known as H9) expressing GFP under the *HES5* promoter⁴ to isolate defined neural progenitor populations of neuroepithelial (NE), early radial glial (ERG), mid radial glial (MRG) and late radial glial (LRG) cells based on their cell morphology and Notch activation state⁵, as well as long-term neural progenitors (LNP) based on their epidermal growth factor receptor (EGFR) expression⁶ (Fig. 1a and Extended Data Fig. 1a). We took these defined stages to create strand-specific RNA sequencing (RNA-seq) data, chromatin immunoprecipitation followed by sequencing (ChIP-seq) maps for histone H3 lysine 4 monomethylation (H3K4me1), trimethylation (H3K4me3), lysine 27 acetylation (H3K27ac) and H3K27me3 as well as DNA methylation (DNAm) data by whole-genome bisulphite sequencing (WGBS) for the first four stages, and reduced representation bisulphite sequencing (RRBS) for the last two (LRG and LNP) stages (Fig. 1a and Supplementary Table 1).

Global transcriptional analysis of the undifferentiated ES cells and the first four neural progenitor cell (NPC) stages identified 3,396 differentially expressed genes (Extended Data Fig. 1b, c and Supplementary Table 2). Pluripotency-associated genes such as *OCT4* (also known as *POU5F1*) and *NANOG* are, as expected, rapidly downregulated, and pan-neural genes are induced early and maintained throughout the remainder of the differentiation time course (Extended Data Fig. 1c).



Using data from the mouse Allen Brain Atlas as an *in vivo* reference for genes expressed in different brain compartments and devel-

opmental stages, we observed a consecutive shift of expression signatures along the NPC differentiation trajectory (Fig. 1b). NE through LRG transcripts suggest anterior neural fates, while the MRG and LRG stages show in addition some posterior identities (Fig. 1b, left). Accordingly, differentiated progeny derived from these populations express deep cortical layer neuronal markers (NEdN and ERGdN) such as *FEZF2* and *BCL11B* and superficial layer neuronal markers (MRGdN) such as *POU3F2/POU3F3* and *MEF2C* (Extended Data Fig. 1d). Progression from early (NE) to late (LRG) stages was also accompanied by a transition from predominantly neurogenic to mainly gliogenic potential, although LRG cells still generate neurons (Extended Data Fig. 1d). This progressive change in NPC identity aligns well with the *in vivo* order of developmental events⁷.

In line with these observations, our WGBS data show changes in DNAm that can be separated into two overall patterns. The first is characterized by widespread loss of methylation and retention of the resulting hypomethylated state throughout subsequent differentiation stages (Fig. 1c, top right). This pattern coincides with major cell-fate decisions such as commitment from ES cells to the neural fate and the transition from ERG to MRG, the latter demarcating both peak of neurogenesis and onset of gliogenic potential (Fig. 1c, right middle). The second pattern is defined by a stage-specific loss with subsequent gain at the next stage, as observed during the transition from NE to ERG and also from MRG to LRG (Fig. 1c, right). Conversely, regions gaining DNAm during transition from one stage to another frequently reside in a hypomethylated state in all preceding stages, indicating the possible silencing of stem cell or pan-neural gene regulatory elements (Fig. 1c, left). At the histone modification level we also observed the most widespread changes during the initial neural induction (Fig. 1d); although it is worth noting that the biggest gain of the repressive mark H3K27me3 occurs at the MRG stage.

These coordinated epigenetic changes are probably the result of differential transcription factor activity^{8–11}. We therefore developed a computational method to attribute the genome-wide changes in histone modifications and DNAm at regions termed footprints to particular transcription factors and quantified this remodelling potential (TERA, transcription factor epigenetic remodelling activity; Fig. 2a, Extended Data Fig. 2a and Methods). Notably, the H3K27ac peak set in our NPC model was significantly enriched for single nucleotide polymorphisms previously reported to be implicated in Alzheimer's disease ($P \leq 0.01$)

¹Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA. ²Harvard Stem Cell Institute, Cambridge, Massachusetts 02138, USA. ³Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, Massachusetts 02138, USA. ⁴Department of Cell and Developmental Biology, Sackler School of Medicine, Tel Aviv University, Ramat Aviv 6997801, Israel. ⁵Department of Pathology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts 02114, USA. ⁶Center for Systems Biology and Center for Cancer Research, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. ⁷Applied Bioinformatics, Center for Bioinformatics and Quantitative Biology Center, University of Tübingen, Tübingen 72076, Germany.

*These authors contributed equally to this work.

§These authors jointly supervised this work.

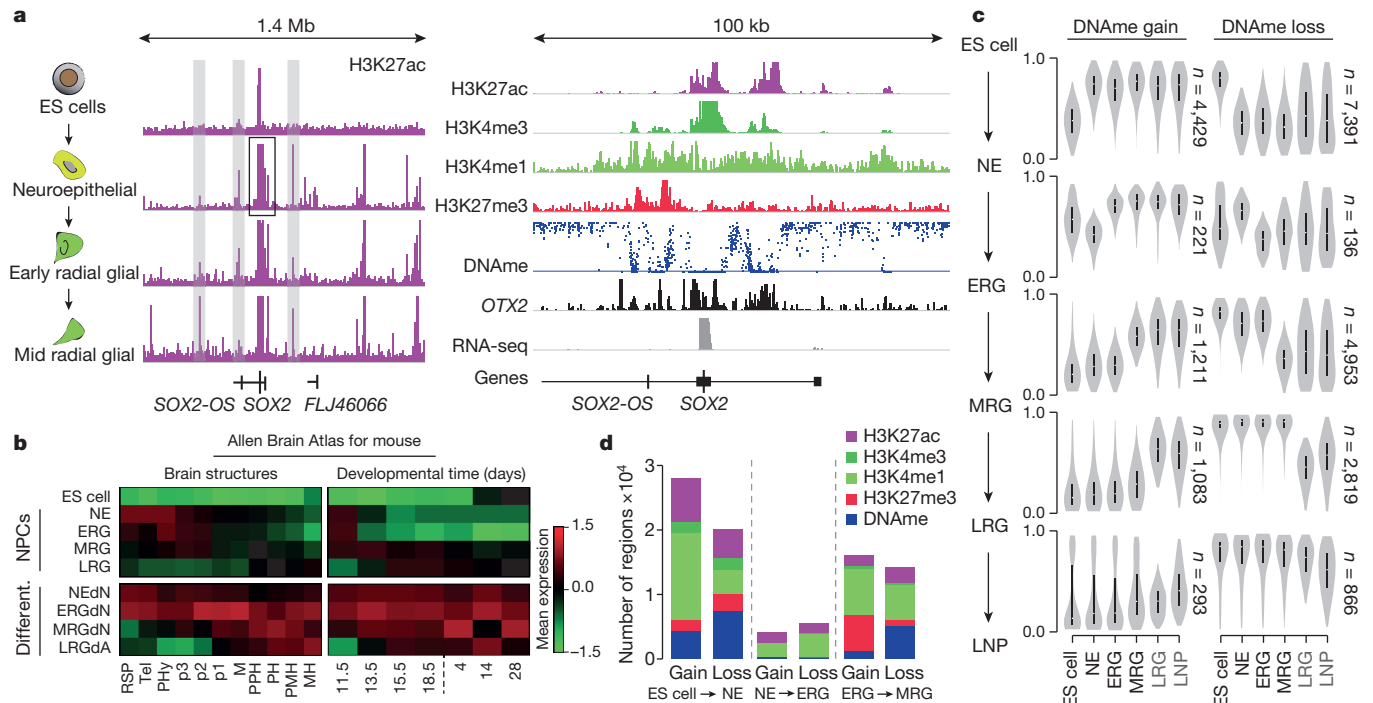


Figure 1 | Consecutive stages of ES-cell-derived neural progenitors are characterized by distinct epigenetic states. **a**, Left, schematic of the cell system. Middle, normalized read-count level for H3K27ac over a 1.4-megabase (Mb) region around the *SOX2* locus (chromosome 3: 180,854,252–182,259,543) where *SOX2*-OS refers to the *SOX2* overlapping transcript. ChIP-seq read counts were normalized to 1 million reads and scaled to the same level (1.5) for all tracks shown. Right, additional tracks for H3K4me3, H3K4me1 and H3K27me3 as well as DNAm (scale 0–100%), *OTX2* binding and expression covering a 100 kilobase (kb) sub-region (chromosome 3: 181,389,523–181,490,148) of this locus. Histone and RNA-seq data were normalized to 1 million reads and are shown on distinct scales. **b**, Maximum gene set activity levels shown as z-scores for genes expressed in defined brain structures (left) and developmental time points (right) based on the mouse Allen Brain Atlas. Gene set activity was defined as average expression level of all member genes followed by z score computation across all nine cell types.

and bipolar disorders ($P \leq 0.01$) by genome-wide association studies, suggesting the possibility to utilize this differentiation system as a basis to study the genetic component of complex diseases *in vitro*^{12,13}. Next, to identify potential key regulators of onset, maintenance and transition through distinct NPC populations, we ranked all motifs and their associated transcription factors based on their TERA scores between consecutive time points (Supplementary Table 3). We then retrieved the transcription factors associated with highest scoring 40 motifs for

Different., differentiated; LRGdA, LRG-derived astrocyte-like cells; RSP, rostral secondary prosencephalon; Tel, telencephalon; PHY, peduncular (caudal) hypothalamus; p3, hypothalamus; p2, pre-thalamus; p1, pre-tectum; M, midbrain; PPH, prepontine hindbrain; PH, pontine hindbrain; PMH, pontomedullary hindbrain; MH, medullary hindbrain. Developmental times are embryonic days 11.5, 13.5, 15.5 and 18.5 and postnatal days 4, 14 and 28. **c**, Distribution of DNAm levels for differentially methylated regions (change in methylation ≥ 0.2 , $P \leq 0.01$) across state transitions; for instance, distributions for regions gaining methylation in the transition from ES cell to NE (top left) at all stages of differentiation. Distinct methylation level trace plots are shown for regions gaining methylation (left) during the specific transitions (indicated on the side) and loss of methylation (right). Black labelled samples are based on WGBS data and grey colour samples (LRG and LNP) were profiled by RRBS. **d**, Bar plot showing the number of regions that gain or lose selected modifications across the first four cell-state transitions.

each cell-state transition (Fig. 2b). This analysis confirmed many well-known key regulators of *in vivo* neural development and forebrain specification that are induced at the NE stage such as *PAX6*, *OTX2* and *FOXG1* (refs 14–16) as well as various SOX proteins¹⁷. Notably, we also found predicted differential activity of distinct downstream components of signalling pathways such as a decrease of *SMAD4* activity at the NE stage, consistent with inhibition of TGF- β signalling that promotes neural induction¹⁸. Another example that is predicted to be relevant

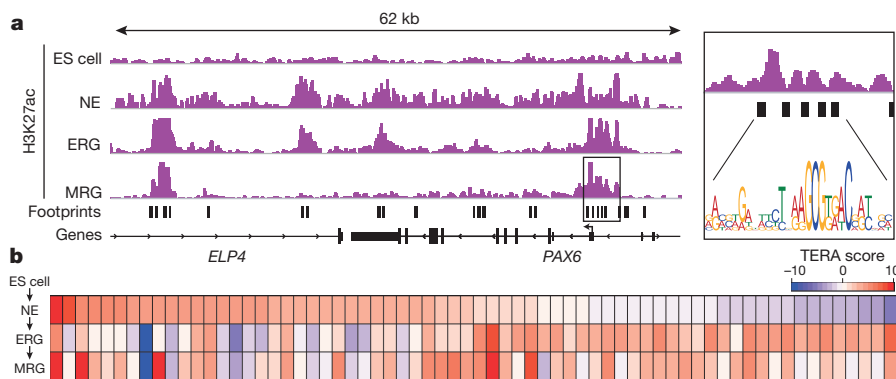


Figure 2 | Distinct transcription factors are associated with stage-specific epigenetic transitions. **a**, Illustration of epigenomic footprinting across the *PAX6* locus (chromosome 11: 31,780,014–31,842,503) for dips in H3K27ac regions (right). Black boxes highlight footprints determined for H3K27ac peaks that harbour various putative transcription factor binding sites based on motif matching. **b**, The 40 top ranked transcription factors predicted to be activated during the cell-state transition are indicated on the bottom. Colour-coding represents normalized transcription factor epigenetic remodelling scores, averaging over all TERAs based on H3K4me3, H3K4me1, H3K27ac and DNAm. In addition, predictions were filtered for factors expressed at the stage of predicted induction.

but not limited to the MRG stage is *POU3F2*, known to be involved in sub-ventricular zone expansion and superficial layer neuronal specification, and *TCF12*, which is highly expressed in germinal zones during brain development¹⁹ (Fig. 2b and Supplementary Table 3).

To obtain a higher-level overview of the processes and roles associated with the distinct putative regulators, we decomposed the H3K27ac data into seven distinct modules, each corresponding to a unique epigenetic dynamic, genomic region and upstream regulator set (Extended Data Fig. 2b, top). Gene set enrichment analysis²⁰ on the genomic regions associated with each of the distinct modules revealed that the module activated upon neural induction and sustained throughout the MRG stage is strongly associated with stem cell maintenance and differentiation-related processes as well as Notch signalling (Extended Data Fig. 2b; module 2). Further analysis of upstream regulators of this module revealed a strong association with *PAX6* and *FOXG1*, suggesting a role for these factors in the general establishment and maintenance of the telencephalic cortical identity of the NPC states (Extended Data Fig. 2c).

To explore the relevance of predicted factors for each cellular state, we carried out a pooled short hairpin RNA (shRNA) screen against 244 transcription factors and epigenetic modifiers selected based on our RNA-seq data (Fig. 3a, Extended Data Fig. 3a and Supplementary Table 4). In total, we recovered 110 factors whose knockdown had a significant (Fig. 3b, q value ≤ 0.05 , mean empirical false discovery rate (FDR) = 0.045, see Methods) negative impact on the number of HES5⁺ cells in at least one differentiation stage (Supplementary Table 4), with high overlap between the distinct stages (Fig. 3c and Extended Data Fig. 3b). Despite the expected high false-negative rate²¹ our screen consistently validated more than 50% of the predicted transcription factors with a known motif for the top 20 motifs found at each stage (Fig. 3d and Extended Data Fig. 3c, d), while an expression-based identification yielded only ~30% recovery (Extended Data Fig. 3c). Among the top factors recovered from the predictions at the early stage (NE and ERG) are the RFX proteins including RFX4, which has been implicated in cortical and brain development^{22,23}, *FOXG1*, as well as *NR2F2*, whose paralogue *NR2F1* has been shown to serve as an intrinsic factor for early regionalization of the neocortex^{24,25}. Gene set enrichment analysis of putative genomic targets of *NR2F2* (see Methods) in the NE cells further expands this role, suggesting involvement in telencephalon, diencephalon and posterior hindbrain development (Supplementary Table 5). At the MRG stage, we recover genes involved in extensive neurogenesis and in commencing early gliogenesis such as *NFLA* and *NFIB*, which are involved in both repressing the neuronal progenitor state through Notch signalling concomitantly with activating glial fates²⁶, as well as *REST*—a major pleiotropic epigenetic regulator of neural cell-fate decisions²⁷.

Next, we selected a set of 22 core factors with evidence to be functional at all stages as assessed by RNA-seq and the shRNA screening results (Extended Data Fig. 4a and Methods). In order to determine whether the subset of core factors with a DNA binding motif available (10 of 22) exerts the same function at each stage, we performed a co-binding analysis based on the predicted binding sites of 523 transcription factors in dynamically regulated distal H3K27ac footprints. This analysis uncovered highly stage-specific relationships that were also supported by the observed knockdown effect at each stage (Fig. 4a and Extended Data Fig. 4b). Notably, most of the identified co-binding partners are either expressed in a more stage-specific fashion or are only activated in more mature neuronal or glial cell types (Fig. 4b). To further validate some of these findings, we focused on *OTX2* due to its high expression in all NPC populations (Fig. 4b) and performed ChIP-seq at the NE and MRG stages. *OTX2* was enriched at more targets in NE cells, of which around 35% overlapped with MRG-bound sites (Fig. 4c and Extended Data Fig. 4c). The shared target set is highly enriched for genes involved in stem cell maintenance and differentiation as well as various pro-neural gene sets known to act during advanced stages of forebrain and midbrain progenitor cell maturation (Fig. 4d and Extended Data Fig. 4d). This binding pattern combined with the observation that the *OTX2* target gene set reaches peak transcriptional activity in the NEdN and ERGdN

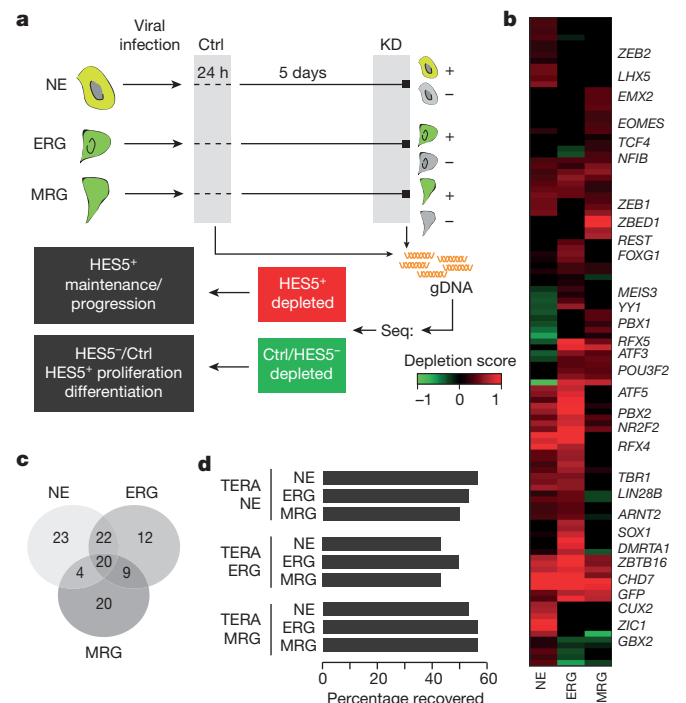


Figure 3 | A pooled shRNA screen recovers predicted regulators of *in vitro* NPC differentiation. **a**, Simplified schematic of the pooled shRNA screen (see Extended Data Fig. 3 for more details). Ctrl, control; gDNA, genomic DNA; KD, knockdown; Seq., sequencing. **b**, Depletion scores for all genes that are significantly reduced (q value ≤ 0.05 for at least two different shRNAs per gene) in at least one stage for fluorescence-activated cell sorting (FACS)-purified HES5⁺ cells 6 days after knockdown compared to FACS sorted HES5⁺ obtained from the same infection or compared to cells collected 24 h after infection (see Extended Data Fig. 3a). Depletion score indicates the extent to which shRNAs targeting a particular gene were lost during the knockdown period relative to the control, indicating potential relevance of a particular gene for HES5⁺ maintenance, NPC state progression and proliferation or cell survival. Higher depletion scores (red) indicate stronger reduction in shRNA presence; scores were capped at 1 and computed based on at least three technical replicates per condition. **c**, Overlap of genes detected to be significantly depleted in the HES5⁺ population relative to at least one of the control conditions. **d**, Performance of combined regulator predictions based on TERA ranking averaged over H3K4me3, H3K4me1, H3K27ac and DNAm. Performance is measured as percentage of the top 20 predicted activating or repressing motifs for each stage mapping to transcription factors included in the shRNA library.

populations implies a role for *OTX2* in the preparation of pro-neural genes expressed at later stages (Fig. 4d, e). These findings further suggest a model where a core set of transcription factors helps sustain NPC identity throughout the differentiation time course and at the same time participates in the progression and modulation of NPC differentiation potential through cooperation with stage-specific regulators.

To gain a better understanding of how factors that are active at distinct NPC stages contribute to their corresponding neuronal and glial cell propensities, we took advantage of the fact that many transcription factor binding sites exhibit a gain of H3K4me1 and loss of DNAm at the early NPC stages before increased expression of their associated genes in more differentiated cell types (hereafter referred to as epigenetic priming) (Fig. 5a and Extended Data Fig. 5a–c). For instance, we identified three pro-neural factors that show evidence of priming, are induced only at a later stage, and possess transcription factor binding sites that are also significantly ($P \leq 0.05$ permutation test) associated with genes differentially expressed at a later stage (Fig. 5a, bold genes). Because these pro-neural genes are not expressed at the early NPC stages but in more mature cell types derived upon mitogen withdrawal, the identification of such priming events highlights that the epigenetic state

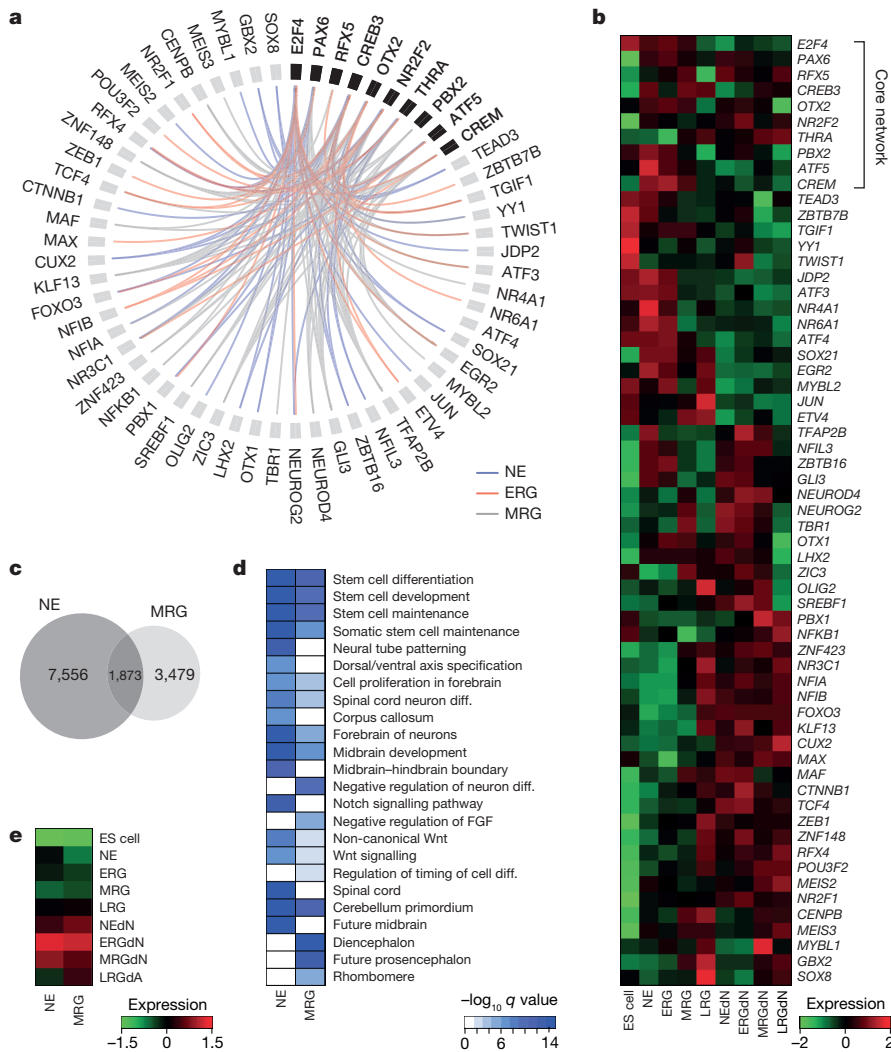


Figure 4 | A set of core transcription factors dynamically associates with stage-specific factors to modulate NPC identity and differentiation potential. **a**, Predicted significant ($P \leq 0.01$, enrichment ≥ 1.5) co-binding relationships in dynamically regulated H3K27ac footprints for a set of 10 transcription factors (bold, core network) required by HES5⁺ cells in at least two stages. Stage-specific predicted co-binding relationships are indicated in blue (NE), red (ERG) and grey (MRG). All predicted relations were filtered for support by a knockdown effect of each gene at the relevant stage. **b**, Gene expression patterns shown as z scores for the core network transcription factors as well as all predicted co-binding partners across ES cells, all NPCs and more mature cellular states. **c**, Venn diagram showing the overlap of OTX2 binding sites determined by ChIP-seq in early NE and MRG cells. **d**, Gene set enrichment analysis results for OTX2 binding sites in early NE and MRG cells. **e**, Median expression patterns for ES cells, all NPCs and more mature cell populations shown as z scores for putative downstream target genes of OTX2 binding sites.

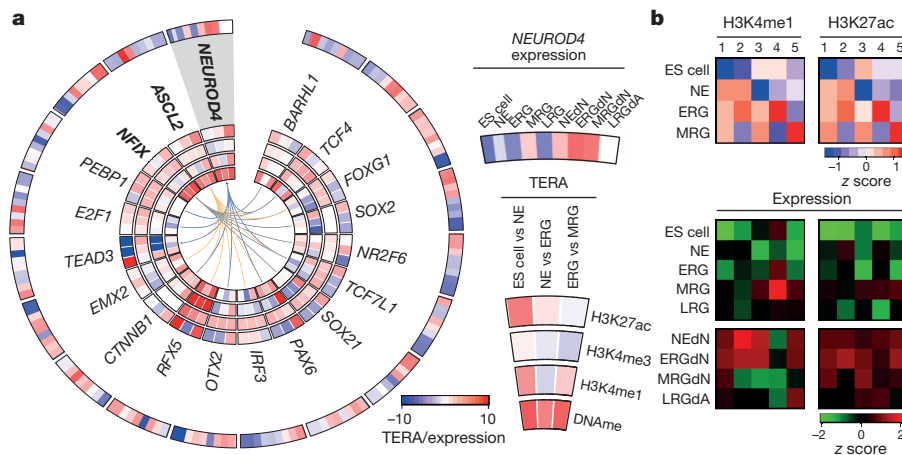


Figure 5 | Binding of core and stage-specific NPC transcription factors is associated with epigenetic priming of pro-neural genes. **a**, Characterization of transcription factors associated with motifs gaining H3K4me1 or losing DNAm at the NE stage before their expression at a later or more differentiated cell state as determined by high TERA scores (bold), termed priming. In addition, significant ($P \leq 0.01$, enrichment ≥ 1.5) co-binding relationships with factors expressed at the NE stage are indicated by coloured lines. For each transcription factor (from outer to inner circles, see example to the right for NEUROD4) heat maps indicating the relative expression level as a z score in all

cell types as well as normalized TERA scores for H3K27ac, H3K4me3, H3K4me1 and DNAm. **b**, Top, heat maps depicting the H3K4me1 (left) and H3K27ac (right) enrichment level for predicted NEUROD binding sites at each NPC stage for five distinct dynamic patterns. At the NE and ERG stages, none of the NEUROD family of proteins is expressed at high levels (< 3.5 fragments per kilobase of transcript per million mapped reads). Bottom, heat map showing the z scores of the median gene expression levels for predicted NEUROD downstream target genes for each of the five dynamic patterns in the more mature neuron- and astrocyte-like populations.

is useful for predicting regulators relevant at later stages of differentiation. In order to pinpoint transcription factors potentially involved in facilitating these priming events at the respective NPC stages, we determined significant predicted co-binding relationships between the subset of pro-neural transcription factors and factors that in contrast are expressed at the stage of priming (Fig. 5a).

To specifically investigate the hypothesis that a part of the pro-neural binding site landscape is epigenetically primed at the NPC stages, we focused on predicted NEUROD protein family binding sites within H3K27ac footprints and defined five patterns of H3K27ac and H3K4me1 enrichments across these sites (Fig. 5b). We found that genes associated with predicted NEUROD binding sites in regions gaining H3K27ac or H3K4me1 enrichment at distinct stages of NPC progression are upregulated in more mature populations derived from the respective NPC stage (Fig. 5b and Extended Data Fig. 5d). Consistent with the idea of a comprehensive preparation of the epigenetic landscape during lineage specification, NEUROD binding sites that retain high levels of H3K27ac and H3K4me1 throughout the entire differentiation time course are associated with various anterior and posterior cortical structures as well as early and late developmental time points (Extended Data Fig. 5e).

These results support a model where selected transcription factors at the NPC stage remodel the binding site repertoire for pro-neural factors by preparing the epigenetic landscape at their respective targets. First the general lineage landscape is established upon commitment to the neural fate, followed by the stage-specific modulation of primed pro-neural binding sites. This in turn might serve as a mechanism to restrict their binding space in order to ensure proper neuronal and glial differentiation capacity. In addition to these insights into the epigenetic dynamics during differentiation, we provide a general analysis strategy to interpret differences in epigenetic landscapes based on cell-fate regulatory transcription factors. This strategy can be readily applied to other data sets including the extensive collection of the NIH Roadmap Epigenomics Project (Supplementary Table 3).

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 19 November 2013; accepted 21 October 2014.

Published online 24 December 2014.

- Imayoshi, I., Sakamoto, M., Yamaguchi, M., Mori, K. & Kageyama, R. Essential roles of Notch signaling in maintenance of neural stem cells in developing and adult brains. *J. Neurosci.* **30**, 3489–3498 (2010).
- Shimojo, H., Ohtsuka, T. & Kageyama, R. Dynamic expression of notch signaling genes in neural stem/progenitor cells. *Front. Neurosci.* **5**, 78 (2011).
- Carlén, M. *et al.* Forebrain ependymal cells are Notch-dependent and generate neuroblasts and astrocytes after stroke. *Nature Neurosci.* **12**, 259–267 (2009).
- Placantonakis, D. G. *et al.* BAC transgenesis in human embryonic stem cells as a novel tool to define the human neural lineage. *Stem Cells* **27**, 521–532 (2009).
- Elkabetz, Y. *et al.* Human ES cell-derived neural rosettes reveal a functionally distinct early neural stem cell stage. *Genes Dev.* **22**, 152–165 (2008).
- Lafaille, F. G. *et al.* Impaired intrinsic immunity to HSV-1 in human iPSC-derived TLR3-deficient CNS cells. *Nature* **491**, 769–773 (2012).
- Lui, J. H. *et al.* Development and evolution of the human neocortex. *Cell* **46**, 18–36 (2011).
- Voss, T. C. & Hager, G. L. Dynamic regulation of transcriptional states by chromatin and transcription factors. *Nature Rev. Genet.* **15**, 69–81 (2014).
- Ziller, M. J. *et al.* Charting a dynamic DNA methylation landscape of the human genome. *Nature* **500**, 477–481 (2013).
- Gifford, C. A. *et al.* Transcriptional and epigenetic dynamics during specification of human embryonic stem cells. *Cell* **153**, 1149–1163 (2013).
- Arnold, P. *et al.* Modeling of epigenome dynamics identifies transcription factors that mediate Polycomb targeting. *Genome Res.* **23**, 60–73 (2012).
- Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
- Claussnitzer, M. *et al.* Leveraging cross-species transcription factor binding site patterns: from diabetes risk loci to disease mechanisms. *Cell* **156**, 343–358 (2014).
- Götz, M., Stoykova, A. & Gruss, P. Pax6 controls radial glia differentiation in the cerebral cortex. *Neuron* **21**, 1031–1044 (1998).
- Hanashima, C., Li, S. C., Shen, L., Lai, E. & Fishell, G. Foxg1 suppresses early cortical cell fate. *Science* **303**, 56–59 (2004).
- Martinez-Barbera, J. P. *et al.* Regionalisation of anterior neuroectoderm and its competence in responding to forebrain and midbrain inducing activities depend on mutual antagonism between OTX2 and GBX2. *Development* **128**, 4789–4800 (2001).
- Pevny, L. H., Sockanathan, S., Placzek, M. & Lovell-Badge, R. A role for SOX1 in neural determination. *Development* **125**, 1967–1978 (1998).
- Chambers, S. M. *et al.* Highly efficient neural conversion of human ES and iPS cells by dual inhibition of SMAD signaling. *Nature Biotechnol.* **27**, 275–280 (2009).
- Uittenbogaard, M. & Chiaramello, A. Expression of the bHLH transcription factor Tcf12 (ME1) gene is linked to the expansion of precursor cell populations during neurogenesis. *Gene Expr. Patterns* **1**, 115–121 (2002).
- McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnol.* **28**, 495–501 (2010).
- Sims, D. *et al.* High-throughput RNA interference screening using pooled shRNA libraries and next generation sequencing. *Genome Biol.* **12**, R104 (2011).
- Blackshear, P. J. *et al.* Graded phenotypic response to partial and complete deficiency of a brain-specific transcript variant of the winged helix transcription factor RFX4. *Development* **130**, 4539–4552 (2003).
- Zarbalis, K. *et al.* A focused and efficient genetic screening strategy in the mouse: identification of mutations that disrupt cortical development. *PLoS Biol.* **2**, e219 (2004).
- Zhou, C., Tsai, S. Y. & Tsai, M. J. COUP-TFI: an intrinsic factor for early regionalization of the neocortex. *Genes Dev.* **15**, 2054–2059 (2001).
- Faedo, A. *et al.* COUP-TFI coordinates cortical patterning, neurogenesis, and laminar fate and modulates MAPK/ERK, AKT, and β -catenin signaling. *Cereb. Cortex* **18**, 2117–2131 (2008).
- Piper, M. *et al.* NFIA controls telencephalic progenitor cell differentiation through repression of the Notch effector *Hes1*. *J. Neurosci.* **30**, 9127–9139 (2010).
- Qureshi, I. A., Gokhan, S. & Mehler, M. F. REST and CoREST are transcriptional and epigenetic regulators of seminal neural fate decisions. *Cell Cycle* **9**, 4477–4486 (2010).

Supplementary Information is available in the online version of the paper.

Acknowledgements We would like to thank all members of the Meissner and Elkabetz laboratories; we thank L. Studer (Sloan-Kettering Institute) for the *HES5::eGFP* reporter line; we also thank F. Kelley and other members of the Broad Sequencing Platform, J. Doench and members of the Genome Perturbation Platform at the Broad Institute, D.-A. Landau for critical reading of the manuscript, as well as to I. Shur and O. Sagi-Assif at Tel Aviv University for their extensive FACS operation. We also thank L. Gaffney for graphical support. This work was funded by the NIH Common Fund (U01ES017155), NHGRI (HG006911), NIGMS (P01GM099117), the New York Stem Cell Foundation, the Israel Science Foundation (ISF) (1126/10, 1710/10) and a Marie Curie International Reintegration Grant (IRG277151). A.G. is supported by the Charles H. Hood Foundation and A.M. is a New York Stem Cell Foundation Robertson Investigator.

Author Contributions The study was designed by M.J.Z., Y.E. and A.M. R.E. and Y.E. developed the NPC system, R.E. and Y.Y. performed consecutive cell isolation, propagation and differentiation and conducted the shRNA screen with Y.E.. M.J.Z. performed the analysis and designed the shRNA screen. W.M. and J.L.R. helped with RNA-seq data processing and analysis. J.D. performed transcription factor ChIP-seq experiments. R.P. and C.A.G. performed RNA-seq library construction. R.P. and D.C. performed shRNA library construction. T.S.M. provided experimental advice. R.I., J.X. and A.G. conducted histone ChIP-seq experiments. H.G. performed WGBS and RRBS library construction. A.G. and A.M. supervised the DNA methylation profiling. C.E. and B.E.B. provided experimental input and advice for the ChIP-seq experiments. A.M.T. provided the transcription factor ChIP-seq protocol. O.K. assisted in the design of analytical methods. M.J.Z., Y.E. and A.M. interpreted the data and wrote the manuscript.

Author Information All data are available from the GEO database under accession number GSE62193, the NIH Roadmap (<http://www.roadmapepigenomics.org/data>) and NCBI Epigenomics portal (<http://www.ncbi.nlm.nih.gov/epigenomics>). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to Y.E. (elkabetz@tauex.tau.ac.il) or A.M. (alexander_meissner@harvard.edu).

METHODS

Culturing undifferentiated human ES cells. *HES5::eGFP* bacterial artificial chromosome transgenic human ES cells (H9; WA9; Wicell) expressing GFP under the *HES5* promoter were cultured on mitotically inactivated mouse embryonic fibroblasts (MEFs) (Globalstem). Undifferentiated ES cells were maintained as described previously⁵ in medium containing DMEM/F12, 20% KSR, 1 mM glutamine, 1% penicillin/streptomycin, non-essential amino acids and β -mercaptoethanol. Undifferentiated ES cells were purified with pluripotency markers Alexa 647-conjugated Tra-1-60 and phycoerythrin-conjugated SSEA-3 (BD Pharmingen).

Neural induction and long-term propagation of NPCs. Neural differentiation of ES cells was performed as described in refs 5,18. In brief, neuroepithelial cells were generated either by monolayer induction—with dissociated ES cells plated on Matrigel (BD Biosciences)—or by co-culture on MS5 stromal cells. In both cases neural fate was directed by dual SMAD inhibition protocol¹⁸. Neural rosettes generated from both induction methods were harvested mechanically during all stages of differentiation and replated on culture dishes pre-coated with $15 \mu\text{g ml}^{-1}$ polyornithine (Sigma), $1 \mu\text{g ml}^{-1}$ laminin (BD Biosciences) and $1 \mu\text{g ml}^{-1}$ fibronectin (BD Biosciences) (Po/Lam/FN) in N2 medium composed of DMEM/F12 and N2 supplement (Invitrogen). N2 supplement contained insulin, *apo*-transferin, sodium selenite, putrescine and progesterone. This medium was supplemented with sonic hedgehog (30 ng ml^{-1}), fibroblast growth factor 8 (FGF8; 100 ng ml^{-1}) and brain-derived neurotrophic factor (BDNF) (20 ng ml^{-1}) (all from R&D Systems) to induce and maintain early anterior regionalization of NE cells. These factors were gradually replaced by FGF2 (20 ng ml^{-1}) and EGF (20 ng ml^{-1}) in the following 2 weeks of differentiation in order to maintain a proliferative (FGF and EGF responsive) NPC state. NPCs from all stages were collected at indicated days and FACS purified for *HES5::eGFP* (NE to LRG) or EGFR for LNPs to purify for the highest NPC state for each stage. NE cells were collected at day 12 of differentiation, ERG cells were collected at day 14, mid-neurogenesis radial glial (MRG) cells were collected at day 35, late-gliogenic radial glial (LRG) cells were collected at day 80, and long-term NPCs (LNP) were collected at day 220. At each stage cells were either split for the next passage or subjected to FACS purification for *HES5::eGFP* as described. All replating was performed on Po/Lam/FN-coated dishes. For generating mature differentiated populations, *HES5*⁺ sorted NPCs were seeded at high density and subjected to mitogen withdrawal differentiation medium for 17 days which included N2 supplemented with ascorbic acid/BDNF (neuronal; NEDN, ERGdN, MRGdN) or 5% fetal bovine serum (FBS) (Invitrogen) (glial; LRGdA). Additional experimental details and in-depth characterization of these cell types are provided in Elkabetz and colleagues (manuscript in preparation).

Chromatin immunoprecipitation followed by sequencing (ChIP-seq). For the histone ChIP experiments, we used similar approaches to ref. 28. Specifically, around 160,000 cells were crosslinked in 1% formaldehyde for 10 min at 37°C , followed by quenching with 125 mM glycine for 5 min at 37°C , washed with PBS containing protease inhibitor (Roche, 04693159001) and flash-frozen in liquid nitrogen. To lyse the cells, we used 1% SDS, 10 mM EDTA and 50 mM Tris-HCl, pH 8.1 complemented with a protease inhibitor. The chromatin was then fragmented with a Branson Sonifier (model S-450D) at 4°C , and calibrated to a size range of 200 and 800 base pairs (bp). Chromatin was mixed with antibody and incubated at 4°C overnight. Protein A and Protein G Dynabeads were added to chromatin/antibody mix (Invitrogen, 100-02D and 100-07D, respectively) and incubated for 1–2 h at 4°C . Samples were washed six times with RIPA buffer (10 mM Tris-HCl, pH 8.0, 1 mM EDTA, pH 8.0, 14 mM NaCl, 1% Triton X-100, 0.1% SDS, 0.1% DOC), twice with RIPA buffer containing 500 mM NaCl, twice with LiCl buffer (10 mM TE, 250 mM LiCl, 0.5% NP-40, 0.5% DOC), twice with TE (10 mM Tris-HCl, pH 8.0, 1 mM EDTA), and then eluted in elution buffer (10 mM Tris-Cl, pH 8.0, 5 mM EDTA, 300 mM NaCl, 0.1% SDS, pH 8.0) at 65°C . Eluate was treated with RNaseA (Roche, 11119915001) and Proteinase K (NEB, P8102S) overnight at 65°C .

For the OTX2 ChIP cells were collected and crosslinked in 1% formaldehyde for 15 min on ice, quenched with 125 mM glycine for 5 min at room temperature and pelleted. Nuclei were then isolated and chromatin was digested at 37°C with MNase enzyme until the majority of the DNA was between 50 and 800 bp. Specifically, 25 U and 35 U of MNase enzyme were used to digest NE cells and RNS/RG cells, respectively. The chromatin was then incubated with the antibodies overnight at 4°C and co-immunoprecipitation of antibody–protein complexes was performed with Protein A or G beads for 1–2 h at 4°C .

All antibody catalogue and lot numbers are listed next to the data set for which they were used in Supplementary Table 1.

ChIP-seq library preparation and sequencing. To extract DNA and create the Illumina libraries we used solid-phase reversible immobilization (SPRI) beads. The SPRI beads were added to the samples, mixed 15 times, and incubated for 2 min at room temperature. Supernatant was extracted from the beads on a magnet (4 min). 70% ethanol was used to wash the beads and then dried for another 4 min. Forty microlitres of EB buffer (10 mM Tris-HCl, pH 8.0) was used to elute

the DNA. The next steps of Illumina library construction include end repair, addition of A-base, ligation of barcoded adaptors and PCR enrichment. To minimize the loss of ChIP material throughout this procedure, we used a general SPRI cleanup procedure after each reaction step reusing the same beads. PEG buffer (20% PEG and 2.5 M NaCl) was used to re-bind ChIP material to SPRI following each reaction, and washing and extraction occurred as stated above. The enzymatic reactions were carried as follows: (1) DNA end-repair: Epicentre End-IT Repair kit incubated at room temperature for 45 min; (2) A-base addition: Klenow ($3'\rightarrow 5'$ exonuclease; New England Biolabs) incubated at 37°C for 30 min; (3) adaptor ligation: DNA ligase (New England Biolabs) and indexed oligo adaptors and incubated at 25°C for 15 min, followed by $0.7\times$ SPRI/reaction to remove non-ligated adaptors; (4) PCR enrichment: PCR mastermix (primer set, dNTP mix, Pfu Ultra Buffer (Agilent), Pfu Ultra-II Fusion (Agilent), water), for 20 cycles. The PCR amplified libraries were cleaned up using $0.7\times$ SPRI/reaction (size selection mode) to remove excessive primers. Roughly 5 picomoles of DNA library was then applied to each lane of the flow cell and sequenced on Illumina HiSeq 2000 sequencers according to standard Illumina protocols.

For the OTX2 ChIP, DNA libraries were constructed using standard Illumina protocols for blunt-ending, poly(A) extension, and ligation. MyOne Silane beads (Life Technologies 37002D) were used to purify DNA fragments following each step of the library preparation. Adaptor ligation was performed overnight at 16°C . Ligated DNA was then PCR amplified and gel size selected for fragments between 150 and 700 bp. Samples were sequenced using Illumina HiSeq at a target sequencing depth of 20 million uniquely aligned reads.

Strand-specific RNA-sequencing library construction. RNA was extracted using the miRNeasy kit (Qiagen, 217004). Poly(A) RNA was isolated using Oligo d (T_{25}) beads (NEB, E7490L). The poly(A) fraction was then fragmented (Invitrogen, AM8740). Fragments smaller than 200 bp were eliminated (Zymo, R1016) and the remaining fraction was treated with FastAP Thermosensitive Alkaline Phosphatase (Thermo Scientific, EF0652) and T4 Polynucleotide Kinase (NEB, M0201L). RNA was then ligated to a RNA adaptor as reporter previously²⁹ using T4 RNA Ligase 1 (NEB, M0204L), which was then used to facilitate complementary DNA synthesis using Affinity Script Multiple Temperature Reverse Transcriptase (Agilent, 600105). More specifically, we used the following adaptors reported in ref. 29: RNA sequencing, RIL-19 3' RNA adaptor: prArGrArUrCrGrGrArArGrGrCrGrUrCrGrUrG/ddC; RNA sequencing, AR17 reverse transcription primer: ACACGACGCTCTCCGA; RNA sequencing, 3Tr3 5' DNA adaptor: pAGATCGGAAGAGCACACGTCTG/ddC; RNA sequencing, PCR enrichment: AATGATACGGCAGCACCCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCAAGCAGAAGACGGCATACGAGATNNNNNNNNGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT.

RNA was then degraded and the cDNA was ligated to a DNA adaptor using T4 RNA Ligase 1 as described previously²⁹. Final library amplification was completed using NEB Next High Fidelity 2X PCT Master Mix (M054L). To clean up the final PCR and removed adaptor dimers, two subsequent $1\times$ and $0.8\times$ SPRI reactions were completed to prepare the final library for sequencing.

Pooled shRNA screen. We selected 244 transcription factors and epigenetic modifiers that were differentially or continuously highly expressed during our *in vitro* differentiation time course in an otherwise unbiased fashion (Supplementary Table 4). In addition, we included GFP, RFP, LacZ and luciferase as internal controls. We then obtained a sub-pool of the human 45K shRNA pool³⁰ distributed by the Broad Institute Genomic Perturbations Platform and the RNAi Consortium (TRC) against these genes. For each gene, five distinct shRNAs were included as well as five scrambled and three empty control vectors, amounting to a total of $1,230 + 8$ shRNAs. The plasmid for shRNA expression under the control of the constitutive U6 shRNA promoter was the lentiviral vector pLKO.1. shRNA pool production and infection conditions were performed as previously described³⁰. Subsequently, we performed calibration experiments to determine to optimal combination of multiplicity of infection (MOI) and puromycin concentration to ensure efficient selection. We identified MOI 0.4 and $1 \mu\text{g ml}^{-1}$ of puromycin as optimal parameters for all stages. We then infected 26 million cells at each stage of NE, ERG and MRG to ensure sufficient shRNA integration events to recover the complexity of the shRNA library. Twenty-four hours post infection and before full expression but after integration of the lentivirus into the genome we collected 3 million cells to determine our baseline shRNA library representation. Subsequently, we subjected the cells to 5 days of puromycin selection and then FACS sorted the resulting populations into *HES5*⁺ and *HES5*[−] compartments. Next, we assessed the representation of the shRNA library in each of the 9 populations by retrieving all shRNA integration events from genomic DNA isolated from each sample using PCR followed by next-generation sequencing as previously described³¹. More specifically, we performed two rounds of PCR using the following primers for the primary PCR: primary reverse: CTTAGTTTGTATGCTGTGTTGCTATTAT; primary forward: AATGGACTATCATATGCTTACCGTAAC. For the second, nested PCR we used: nested

forward: GGCTTTATATATCTTGTGGAAAGGA; nested reverse: GGATGAA TACTGCCATTGTCTC.

Next, we performed standard Illumina sequencing library construction as outlined above for four technical replicates for NE and MRG and three technical replicates for ERG, each comprising HES5⁺, HES5⁻ and 24-h control, amounting to a total of 33 libraries. We then sequenced these amplicon libraries on a HiSeq2500 with a PhiX spike-in of 25%.

Individual shRNA validation for OTX2 and PAX6. RNA was extracted using miRNeasy kit (Qiagen) followed by Maxima reverse transcription reaction kit (Fermentas). One nanogram of cDNA was subjected to quantitative PCR (qPCR) using our custom-designed primers and the ABsolute QPCR SYBR Green ROX Mix (ABgene) on a ViiA-7 cycler (ABI). Threshold cycle values were determined in triplicates and presented as average compared to HPRT. Fold changes were calculated using the $2^{-\Delta\Delta C_T}$ method.

WGBS and RRBS library production. WGBS libraries were generated as previously described in ref. 10. RRBS was carried out using the multiplexed, gel-free protocol described in ref. 32.

Data processing. For RNA-seq data processing, reads were trimmed to 80, 60 or 30 bp depending on their per-base quality distribution to achieve maximum alignment rates. Reads were mapped to the human genome (hg19) using TopHat v2.0 (ref. 33) (<http://tophat.cbcb.umd.edu>) employing the unfiltered gencode.v19. annotation.gtf annotation as the transcriptome reference. TopHat was run with default parameters except for the coverage search being turned off. Transcript expression was estimated with Cuffdiff 2 (ref. 34). The workflow used to analyse the data are described in detail in ref. 35 (alternate protocol B).

WGBS libraries were aligned using BSMAP 2.7 (ref. 36) to the hg19/GRCh37 reference assembly. Subsequently, CpG methylation calls were made using custom software as previously described⁹, excluding duplicate, low-quality reads as well as reads with more than 10% mismatches. Only CpGs with more than 5× coverage were considered for further analysis.

ChIP-seq data were aligned to the hg19/GRCh37 reference genome using MAQ³⁷ version 0.7.1 with default parameter settings or Bowtie 2 version 2.05 (ref. 38). Reads were filtered for duplicates and extended by 200 bp at the end of the read. Visualization of read count data was performed by converting raw BAM files to .tbf files using IGV tools³⁹ and normalizing to 1 million reads. Fragment-length-extended, duplicate and quality-filtered reads were used for subsequent analysis.

shRNA screen data analysis. For the screen data analysis, we followed the protocol outlined in ref. 40 employing the R package limma⁴¹. First, we extracted and counted the number of times each shRNA was observed in each library using the shRNA sequence as barcode and the R function processHairpinReads(). Next, we normalized the shRNA counts to the total number of reads observed containing a shRNA to counts per million (cpm) and retained only those shRNAs with more than 0.5 cpm in more than 2 samples. After further quality control showing excellent reproducibility (Extended Data Fig. 3f), we performed differential shRNA count analysis between the HES5⁺ and 24-h control and the HES5⁺ and HES5⁻ populations for each stage. To that end we first estimated the dispersion for each condition and then fitted a negative binomial generalized linear model using the R package edgeR. We then conducted a likelihood ratio test for each contrast and only retain those shRNAs as differentially enriched at a FDR ≤ 0.05. To determine genes with significant positive or negative impact on HES5⁺ maintenance or cell survival, we determined all genes that were targeted by at least two independent shRNAs which showed a significant effect (FDR ≤ 0.05) in the same direction. We then computed a mean effect score in order to rank genes by computing the weighted mean of the log fold change between the two conditions weighted by the log cpm across all significant shRNAs and targeting a particular gene with an effect in the same direction. If an equal number of shRNAs showed a significant effect in positive or negative direction, we classified the gene as not significantly affected. Otherwise we chose the effect direction based on the majority of the shRNAs. We then combined the results from the HES5⁺ to 24-h control and HES5⁺ comparison into one by taking the maximum mean effect score observed in either comparison. The resulting mean effect scores are then used for visualization and analysis purposes in main text and figures and are reported in Supplementary Table 3. In addition, we also calculated an empirical FDR by determining the fraction of shRNAs with a statistically significant effect based on the generalized linear model but were not expressed based on the RNA-seq data for the condition where the significant effect was observed.

For the TERA validation analysis, we ranked all motifs according to their TERA scores at each stage. Next, we filtered out motifs that were not associated with at least one transcription factor that was covered in our screen design. We then determined the fraction of top 20 motifs (by absolute TERA values) that were linked to transcription factors which showed a significant effect in the corresponding stage-specific shRNA screen. We report this number as the percentage of motifs recovered. Only motif-knockdown results that have a straightforward interpretation

were considered as hits. These include: (1) positive TERA score and positive depletion score (gene is involved HES5⁺ maintenance, progression or cell survival); (2) negative TERA score and negative depletion score (impedes HES5⁺ maintenance, progression or apoptosis); (3) negative TERA score and positive depletion score (gene is involved HES5⁺ maintenance, progression or cell survival but most likely acts as a repressor by causing H3K27ac or H3K4me3/1 loss). For the comparison with the expression-based analysis, we ranked all significantly differentially expressed genes by their absolute fold change and determined the fraction of top 20 transcription factors observed among the differentially enriched shRNAs in the screen.

Differential expression analysis. Differential expression analysis was carried out using Cuffdiff 2 (ref. 34) and genes differentially expressed at a FDR ≤ 0.1 for each comparison and a minimal expression level of 1 FPKM in at least one of the conditions were considered. Clustering analysis was performed using the cCluster() function in the cummeRbund⁴² package version 2.6.1 (<http://compbio.mit.edu/cummeRbund/>) with the Jensen–Shannon distance as metric. The number of clusters for the NPC set (ESC, NE, ERG, MRG, LRG) and the differentiated populations (NEdN, ERGdN, MRGdN, LRGdA) was determined as the number of clusters between 10 and 20 with the minimum average silhouette width across all clusters. Subsequently, a pseudocount of 1 was added to all FPKM counts followed by a log₂ transformation. The resulting values were used for all further expression analysis.

ChIP-seq data analysis and normalization. For H3K27ac and H3K4me3 histone marks, the irreproducible discovery rate (IDR) framework⁴³ with a cutoff of 0.1 in combination with the MACS2 (ref. 44) peak caller version 2.1 was used to identify peaks taking advantage of both replicates for each condition. For MACS2 peak calling, we used an initial *P* value cutoff of 0.01 and the corresponding whole-cell extract (WCE) control library as background. All IDR peak sets can be obtained from GEO under GSE62193.

For the broad histone marks H3K27me3 and H3K4me1, we first determined all 1-kilobase (kb) tiles of the human genome (hg19) that were significantly enriched over background in at least one of the replicates. To that end we used a Poisson model⁴⁵ with the WCE as background to model the fragment count distribution in each genomic To that end we defined a nominal *P* value for enrichment within a given region *i* in sample *k* harbouring r_{ik} ChIP fragments compared to the WCE control sample *l* with r_{il} ChIP fragments as $P(C \geq r_{ik})$ where⁴⁵:

$$C \sim \text{Poisson}(\max[1, e_{il}] \lambda_k)$$

and $e_{il} = r_{il} / \lambda_l$, $\lambda_k = (\text{region size}) \times (\text{total number of ChIP fragments in sample } k) / (\text{corrected genome size})$, $\lambda_l = (\text{region size}) \times (\text{total number of ChIP fragments in sample } l) / (\text{corrected genome size})$. In order to account for regions with no or minimal WCE read counts due to sampling, we chose $e_{il} = \max(e_{il}, 1)$. Resulting *P* values were adjusted for multiple testing using the Benjamini–Hochberg⁴⁶ correction and the *q* value R package⁴⁷. Only regions significant at a *q* value ≤ 0.05 and with an enrichment level over background ≥ 1.5 were considered to be enriched.

For differential enrichment analysis of histone marks between consecutive conditions, we used the R package diffBind⁴⁸. To normalize read counts, we used the effective library size, counting only reads in peak regions (either the IDR peaks for H3K27ac, H3K4me3 or the enriched 1-kb tiles for H3K27me3 or H3K4me1). The differential analysis was then conducted using the DBA_DESEQ2 method, taking full advantage of both replicates per condition with the bTagwise parameter set to true. Only regions that were differentially enriched between consecutive conditions at a *P* value of 0.05 were reported.

In addition, we created a union peak set for each mark separately by joining overlapping peaks/enriched regions in preparation for the TERA analysis. For H3K4me1, we computed the enrichment over the union of all H3K27ac regions since we wanted to focus on much more sharply defined putative enhancer regions for this mark. For H3K27ac, we focused on distal regions only (≥ 1 kb from nearest TSS) since we were specifically interested in putative enhancer regions for this mark. For H3K4me3, we used the union of all H3K4me3 IDR based peaks regardless of distance, accounting for most promoters and CpG islands. We then determined the enrichment level for all regions in the union set in each replicate across all marks separately. Region enrichment was computed as follows: first, the number of tag counts in each region was determined and normalized to reads per kilobase per million reads (RPKM) sequenced using the full library size of non-duplicate reads. Next, RPKM read counts were divided by the mean RPKM counts across all WCE libraries. Subsequently, the resulting enrichment levels were log₂ transformed. Finally, the resulting enrichment values were quantile normalized across the entire data set for each mark separately. The resulting values were then average across replicates to obtain a region × condition normalized enrichment matrix. The resulting matrix was used as input for the TERA analysis. We tested several ChIP normalization strategies by assessing between-replicate correlation

and between-condition discriminative power on a large data set of 70 REMC H3K27ac samples and identified this strategy as the best performing one.

Footprinting detection. To determine small regions depleted of histone modifications but surrounded by regions of much greater enrichment, termed footprints, we extended an approach used for the analysis of DNase I hypersensitivity (HS) data⁴⁹. Our footprints identification algorithm consisted of three main phases. In the first phase, we identify peaks using the IDR framework (see previous section) for H3K27ac and H3K4me3 and use these as baseline regions in which footprints could be detected. In the second phase, we identified footprints located within/around peak regions in the following manner. (1) For each peak, extend by 400 bp from apex in either direction. (2) Split entire resulting region into bins of size 20 bp. (3) Compute number of RPKM counts for a central sliding window across the entire region (shifting by increments of one bin) for different window sizes ranging from two bins to ten bins in increments of one. (4) For each position of the central window and for each window size, compute the following three quantities: C_{ij} – RPKM count for central window at current position i and window size j , R_{ij} – RPKM count for a 200-bp stretch directly to the right of the central window and L_{ij} – RPKM count for a 200-bp stretch directly to the left of the central window. (5) For each resulting position i and window size j compute the depletion score:

$$e_{ij} = \frac{f(C_{ij}+1)}{2L_{ij}} + \frac{f(C_{ij}+1)}{2R_{ij}}$$

With the footprint size normalization factor $f = s/b$, with s the size of the central window and b the size of the border regions. (6) Identify non-overlapping, non-adjacent footprint candidates starting from small to larger central window sizes and recording footprint candidate if $e_{ij} > 0$ and $e_{ij} < 1$ and $L_{ij} > C_{ij}$ and $R_{ij} > C_{ij}$, followed by removing all other potential footprints (central window + borders) of larger size overlapping the current candidate. (7) Finally, all resulting candidate footprints with a footprinting score $e_{ij} \leq 0.9$ were reported.

The latter procedure was carried out for H3K27ac and H3K4me3 independently for each sample. Subsequently, we merged all footprints from individual samples into consensus footprints set for each epigenetic mark separately, collapsing overlapping footprints by taking the union of all regions with non-zero overlap.

Differentially methylated region detection. Differentially methylated region (DMR) detection was carried out as previously described with slight modifications¹⁰. Pairwise comparisons of consecutive samples (hESC, NE, ERG, MRG, LRG, LNP) were carried out on a single CpG level using a β -binomial model and the β difference distribution requiring a maximum q value below 0.05 and an absolute methylation difference greater than 0.1. q values were computed based on β -binomial model P values using the Benjamini–Hochberg⁴⁶ method. Only CpGs covered by at least 5 reads in either sample were considered. Subsequently, differentially methylated CpGs within 500 bp were merged into discrete regions. Differential CpGs without neighbours were embedded into a 100-bp region surrounding each CpG. Next, differential methylation analysis was repeated on the region level using a random effects model. Only regions significant at a P value below 0.01, an absolute methylation difference above 0.2 and containing at least 2 differentially methylated CpGs were considered differentially methylated. These regions were defined as DMRs and used for subsequent analysis. For the DNA methylation analysis in the context of the TERA framework, we restricted our analysis to DMRs consistently covered across all conditions, including those only assessed by RRBS. This left us with 7,929 regions.

Association of genomic regions with genes. We used the R package ChIPpeakAnno⁵⁰ to associate each region with its nearest ENSEMBL transcription start site and used this mapping for all downstream analysis.

Gene set enrichment analysis. Gene set enrichment analysis for genomic regions was carried out using the GREAT toolbox²⁰ and only categories with q values ≤ 0.05 for both the hypergeometric and the binomial test as well as a minimal region enrichment level greater than 2 were considered, following the GREAT recommendations. Due to the large number of enriched gene sets, a selected subset of the results is shown in the different figures. In addition, we used the Allen Brain Atlas⁵¹ to determine enrichment for distinct brain structures and developmental time points. To that end we derived gene sets from the brain atlas data in the following fashion.

We obtained *in situ* hybridization counts for the developing mouse brain at 7 distinct fetal time points and 11 different brain substructures through direct correspondence with <http://www.alleninstitute.org>. Specifically, we investigated the following structures: rostral secondary prosencephalon (RSP), telencephalon (Tel), peduncular (caudal) hypothalamus (PHy), hypothalamus (p3), pre-thalamus (p2), pre-tectum (p1), midbrain (M), prepontine hindbrain (PPH), pontine hindbrain (PH), pontomedullary hindbrain (PMH), medullary hindbrain (MH); and time points: embryonic (E) day 11.5, E13.5, E15.5 and E18.5 as well postnatal (P) P4, P14 and P28. In total, we had 14,585 measurements for 2,105 different genes

across these different regions and time points. In order to define sets of genes characteristic for each combination of time point and structure, we computed the z scores as well as the maximum observed variation for each gene across the entire matrix of structure and developmental time point combinations. Only genes that exhibited a maximum observed variation (maximum activity – minimum activity) ≥ 1 were considered for gene set definition. Next, we mapped all mouse genes to their human orthologues using the biomaRt database. Finally, we defined gene sets for each region–time–point combination using genes that exhibited a z score ≥ 2 in that particular combination. Since the Allen Brain Atlas gene sets are defined for each developmental time point and regional identity, we next simplified the visualization by focusing either exclusively on structures or developmental time points. Therefore, we determined the gene set with the maximum gene set activity at each differentiation stage across all gene sets associated with distinct developmental time points for each structure separately. Similarly, we determined the gene set with maximum activity for each developmental time point now taking the maximum across all structures at each stage. The gene set activity was determined as the mean \log_2 -transformed expression level of all gene set members in for each condition.

Motif library construction and mapping to transcription factors. We combined the position weight matrices (PWM) from Transfac professional database⁵² (2011) with the PWM collection reported in ref. 53, only retaining motifs annotated for *Homo sapiens* or mouse. To eliminate redundant motifs, we determined pairwise motif similarities for all resulting 1,886 PWMs using the TOMTOM⁵⁴ program which is part of the MEME⁵⁵ suite with default parameters. Next, we compiled a pseudo-distance matrix based on the resulting pairwise motif similarities. As a proxy for motif similarity, we used the \log_{10} -transformed TOMTOM q value which was capped at ten. To convert the resulting motif similarities into a distance matrix, we inverted the scale by subtracting the transformed q values from ten. We then used the resulting matrix to perform hierarchical clustering with Euclidean distance and Ward's method. Finally, we employed the `cutree()` function with a threshold of seven to partition the resulting clustering dendrogram into discrete clusters of motifs. For each cluster, we then determined the motif with the highest complexity based on the relative entropy compared to a genome background model with the following base frequencies: A = 0.2725, C = 0.189, G = 0.189 and T = 0.2728. Only motifs with a relative entropy greater than or equal to eight were retained for subsequent analysis. After identification of the candidate with the highest complexity for each motif cluster, we assigned all genes mapping to any motif in each corresponding cluster to the cluster representative motif. This led to a final motif list of 557 motifs. To obtain a more quantitative association of each motif with its linked genes, we computed the epigenetic transcription factor activity (ETFA) scores across 70 REMC H3K27ac or H3K4me3 cell types and correlated the results with RNA-seq expression data across 40 cell types. This analysis gave rise to a correlation matrix containing the Pearson correlation coefficient of each motif with its linked genes. This matrix was used in combination with the plain gene mapping reported in primary motif sources. For Fig. 2b, we uniquely map each motif to a corresponding linked gene by computing an association score as the product of the absolute Pearson correlation coefficient and the average gene expression level of the corresponding gene. We then chose the gene with the highest association score. For motifs without an entry in the H3K27ac correlation matrix (due to the inability to determine suitable GEV parameters on the REMC data set), we chose the gene with the highest gene expression level. In Fig. 2b, only genes expressed with at least 10 FKPM in the respective condition are considered. We then report the genes mapping to the 40 motifs for each condition, where TERA scores of motifs mapping the same gene were averaged.

In Figs 4 and 5, we incorporated the results of the shRNA screen to uniquely map motifs applying the aforementioned mapping strategy only on the genes identified as hits. If it did not map to any gene hit by the screen, we used the standard assignment strategy outlined above.

Identification of putative transcription factor binding sites. To determine putative binding sites in a given genomic region, we used a biophysical model of transcription factor affinities to DNA^{56,57} to determine putative binding to our footprint sets. This biophysical model requires the training of generalized extreme value (GEV) distributions of binding affinities based on a PWM matrix for each transcription factor and each set of genomic regions in order to generate a suitable background model. In order to take the distinct properties of footprints determined from different epigenetic marks into account, we determined the GEV parameters for footprints arising from H3K27ac, H3K4me3 and DNase using the framework outlined in refs 56, 57. The resulting three binding matrices were then filtered for minimal significant binding affinity at P values below 0.05. All other entries with higher P values were set to one. Next, we took the negative \log_{10} of the entire matrix as a quantitative measure of binding affinity in subsequent analysis.

Inference of transcription factor activities based on epigenetic data. To infer transcription factor epigenetic remodelling activities (TERA), we first computed

ETFA from our epigenetic data. To that end, we first focused on motif activity analysis and associated each motif in a second step with its corresponding transcription factor. For each epigenetic mark, we used the normalized epigenetic enrichment scores as well as DMRs with a minimal DNA methylation difference of at least 0.2 and covered consistently in all data sets. For the DNA methylation data, we inverted the scale to obtain demethylation scores (1 = fully demethylated, 0 = fully methylated) since usually the demethylated states coincides with gene regulatory element activity. To determine the unobserved activity of a transcription factor binding motif, we took advantage of recent developments in the microarray field^{58,59} and adapted this approach to epigenetic data. To that end we modelled the enrichment level y_{it} of a particular epigenetic mark at genomic region i and time point t as a linear function of the unknown transcription factor activities. Considering p predictor variables (epigenetic motif/transcription factor activities) and k time points we describe the unknown transcription factor activities X as a $p \times k$ matrix. Incorporating all regions n meeting the above listed criteria, we employ the linear model $Y = A + BX + E$ with the observed matrix of epigenetic enrichment scores Y ($n \times k$), a constant offset matrix A ($n \times k$), the connectivity matrix B ($n \times p$), describing the filtered binding affinities for all transcription factor motifs to all regions and an error term matrix E . Subsequently, we followed the approach outlined in ref. 58 and applied partial least square (PLS) regression and specifically the SIMPLS algorithm⁶⁰ to determine the unknown transcription factor motif activities. The idea in PLS is to employ a linear dimensionality reduction $T = BR$, where the p predictors in X are mapped onto $c \leq \text{rank}(X) \leq \min(p, n)$ latent components T ($n \times c$ matrix), and to compute the weight matrix R not only based on the data matrix B but explicitly taking into account the response matrix Y . The latter strategy maximizes predictive power even for a small number of latent components.

In order to determine the number of latent components for each epigenetic mark and genomic context, we performed cross validation by randomly partitioning the data set 20 times into two-thirds training and one-third test sets. We then chose the number of components such that it minimized the prediction error. The corresponding analysis methodology was implemented in the statistical programming language R adapting the implementation provided in ref. 58. To assess the significance of the resulting ETFA scores, we performed a permutation test by randomly permuting the epigenetic enrichment scores for each gene regulatory element and recomputed the ETFA values on the permuted values. This process is repeated 100 times. Positive ETFA scores are considered to be insignificant and set to 0 if a greater ETFA score is observed more than once on the randomly permuted set and vice versa for negative ETFA scores.

Finally, we determined the TERA scores by computing the differential ETFA scores between consecutive conditions. These scores were determined by subtracting ETFA scores of consecutive time points from each other. Subsequently, we assessed the significance of this difference using a permutation test by randomly permuting the epigenetic enrichment scores across all regions, re-computing the ETFA scores for each conditions and assessing the TERA score between consecutive conditions for each motif. Positive TERA scores are considered to be insignificant and set to 0 if a greater TERA score is observed more than once on the randomly permuted set and vice versa for negative TERA scores.

Co-binding analysis. Co-binding relationships were evaluated using an empirical approach with the entire set of footprints for each epigenetic mark as background. For a given factor i , we determined the footprints set F_i relevant for the current comparison (for example, changing their epigenetic state in particular cell state transition) that were predicted to contain a transcription factor binding site based on the binding model outlined above. Next, we computed the frequency of motif co-occurrence S_{ij}^G across F_i for all other motifs j in our database. To generate a proper null distribution, we randomly sampled $K = 100$ standardized footprint sets G_k each of size $|F_i|$ from the entire footprint collection for the epigenetic mark under study and computed the same test statistic $S_{ij}^{G_k}$ on these sets. Finally, we determined an empirical P value and enrichment over the control based on these quantities by counting the number of instances for which $S_{ij}^{G_k} \geq S_{ij}^G$:

$$P_{ij} = \frac{\left(\sum_k S_{ij}^{G_k} \geq S_{ij}^G \right)}{K}$$

Only co-binding relationships significant at P values ≤ 0.01 , a median enrichment over the control ≥ 1.5 and an expression level ≥ 2 FPKM in at least one condition were retained. For the core factor co-binding analysis, the predicted co-binding relationships were additionally filtered for support by the knockdown data at the stage of predicted co-binding

Validation analysis on ENCODE data. To validate the outlined strategy *in silico* we took advantage of publically available transcription factor ChIP-seq data in four cell lines from the ENCODE⁶¹ project as well as H3K27ac and RNA-seq data for 70 cell types from the REMC project. We downloaded H3K27ac data as well as

processed transcription factor binding data from the ENCODE project for the cell line K562 since abundant transcription factor binding data based on ChIP-seq was available. In addition, this data set has been successfully used in several studies to benchmark transcription factor binding predictions^{62,63}. We then applied our TERA pipeline to the H3K27ac data sets and computed the transcription factor binding affinities for a set of 557 distinct motifs. With these data sets at hand, we computed the true-positive rate (TPR), the false-positive rate (FPR) and the positive predictive values (PPV) for all transcription factors that could be matched to at least one motif with available binding affinities (46 out of 117). In the event that one factor matched multiple motifs, we chose the motif with the highest area under the curve.

GWAS analysis. The GWAS analysis was conducted using 11,027 GWAS SNPs from the GWAS catalogue (August 2013). We sought to determine whether the H3K27ac-positive regions identified in the NPC populations were enriched for any GWAS SNP class with respect to a H3K27ac peak compendium across many different tissues. To determine a proper background distribution we randomly sampled $K = 1000$ equally sized peak sets from H3K27ac-based footprints identified across 70 epigenome roadmap data sets. Prior to further analysis, we normalized the size of each peak all sets by extending it by 250 bp in each direction from the center coordinate. Next, we determined the overlap with GWAS SNPs for control and neural H3K27ac footprint sets. Subsequently, we computed an empirical P value for each trait/disease i in the catalogue by determining the number of trait associated SNPs S_{ij}^C overlapping with each control region set C_j and the number overlapping with the corresponding footprint set s_i according to

$$P_i = \frac{\left(\sum_j s_i \geq S_{ij}^C \right)}{K}$$

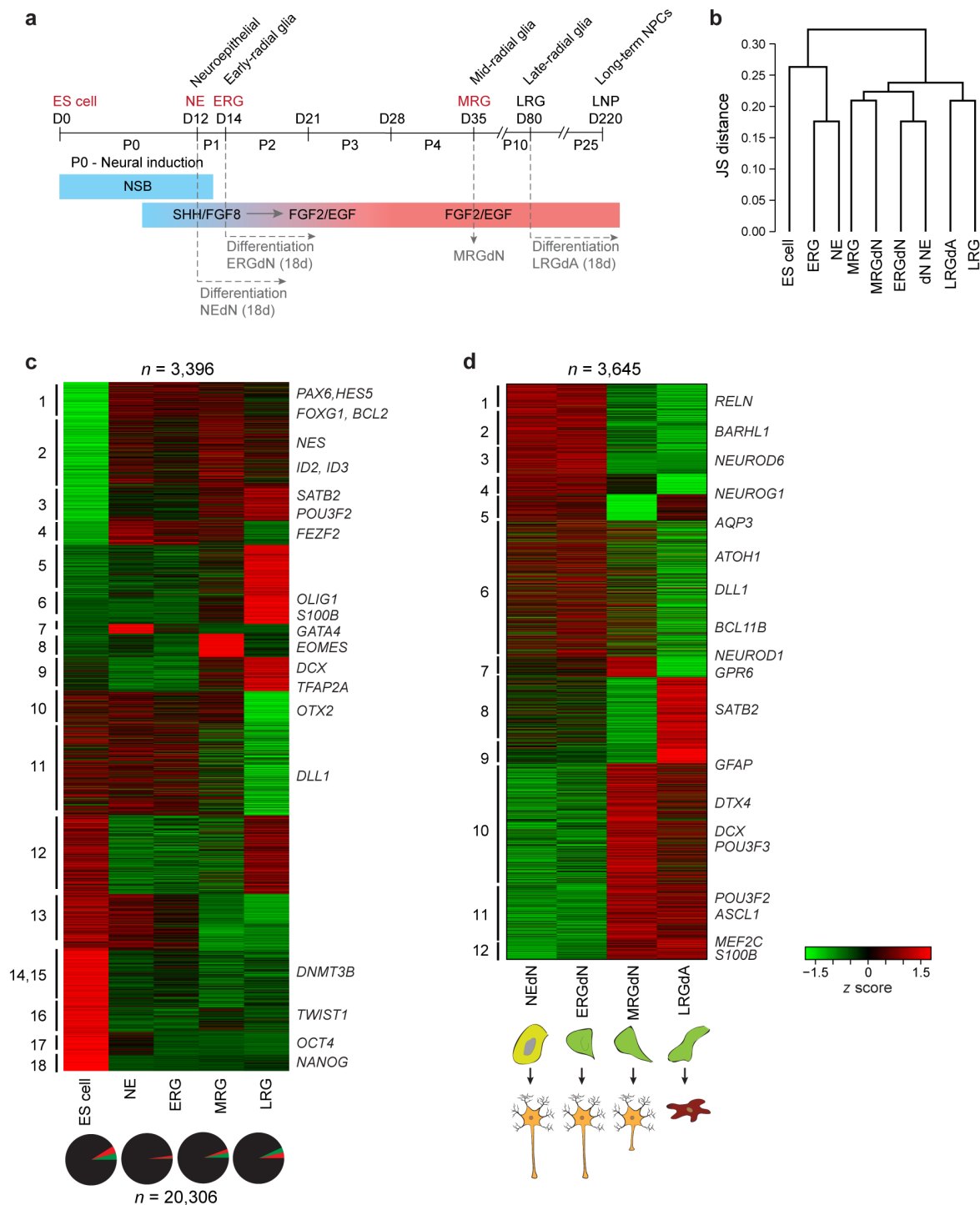
Determination of core network. The core network was defined as those transcription factors that were differentially expressed during neural induction from ES cell to NE and not differentially expressed between consecutive stages of NE, ERG and MRG. We did not consider the LRG stage. Furthermore, we required that each factor was expressed at least 10 FPKM or more in NE, ERG and MRG and that its mean normalized, maximum difference in expression levels between any of the stages did not exceed one standard deviation computed across the entire data set of 9 cell types. In addition, we also considered genes that were not differentially expressed between any consecutive stages including the ESC stage but fulfilled all other criteria. This identification procedure gave rise to the candidate list of core factors. We then intersected this list with the results of our shRNA screen and retained only those factors that were significantly depleted in the HES5⁺ population relative to the respective HES5⁻ or control population in at least two stages. Since the literature supported a role for PAX6 and OTX2 for which our shRNAs showed no effect due to the pooled setup or absent knockdown (Fig. 3f and Extended Data Fig. 3g), we included these genes as well. Finally, we merged this list with all transcription factors that were depleted in our shRNA screen at all three stages in the HES5⁺ population relative to the controls and were expressed at least at 10 FPKM or more in NE, ERG and MRG. This algorithm yielded a list of 22 transcription factors or epigenetic modifiers (Fig. 4a). We then carried out co-binding analysis in H3K27ac footprints dynamically regulated at each stage in order to obtain putative stage-specific co-binding relationships. To determine significant co-binding events, we used the permutation procedure outlined above and retained all co-binding partners with an enrichment over the control ≥ 1.5 that were significant at $P \leq 0.01$ that were also identified as a significant hit in the shRNA screen at the particular stage under investigation.

Transcription factor binding site priming analysis. To determine transcription factors associated with transcription factor binding site priming before factor activation, we determined all transcription factors at each stage that were significantly upregulated at the consecutive NPC time point or induced in the corresponding more differentiated cell type (q value ≤ 0.1) and showed an increase in H3K4me1- or DNase-derived TERA activity at the current stage under investigation. In addition, we required that the corresponding motif did not map to any transcription factor that was expressed more than 3.5 FPKM at the current stage under investigation. From this list, we picked the pro-neural genes *NEUROD4*, *ASCL2* and *NFIX* for further investigation due to their literature support for their pro-neural functions. Finally, we required that the potential downstream target genes were significantly enriched for differentially regulated genes at the next NPC stage or in the corresponding more differentiated cell types. To that end, we determined all putative transcription factor binding sites for a particular factor in dynamically regulated H3K27ac or H3K4me1 footprints at the stage of potential priming. We then associated each of these putative binding sites with the nearest TSS and determined the number of differentially expressed genes for each factor. To assess significance, we randomly drew 100 sets of equally sized H3K27ac footprints with no motif of the factor under investigation and determined the

number of differentially expressed genes for the subsequent stages. Only factors that exhibited more differentially expressed genes compared to the control sets in more than 99% of the cases were retained.

Next, we performed co-binding analysis in H3K27ac peaks differentially regulated between the ES cell and NE stage as outlined above and display the top 10 co-binding relationships per factor with an odds-ratio ≥ 1.5 that were significant at a permutation-test-based $P \leq 0.01$ in Fig. 5a.

28. Garber, M. *et al.* A high-throughput chromatin immunoprecipitation approach reveals principles of dynamic gene regulation in mammals. *Mol. Cell* **47**, 810–822 (2012).
29. Engreitz, J. M. *et al.* The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. *Science* **341**, 1237973 (2013).
30. Luo, B. *et al.* Highly parallel identification of essential genes in cancer cells. *Proc. Natl Acad. Sci. USA* **105**, 20380–20385 (2008).
31. Strezoska, Z. *et al.* Optimized PCR conditions and increased shRNA fold representation improve reproducibility of pooled shRNA screens. *PLoS ONE* **7**, e42341 (2012).
32. Boyle, P. *et al.* Gel-free multiplexed reduced representation bisulfite sequencing for large-scale DNA methylation profiling. *Genome Biol.* **13**, R92 (2012).
33. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-seq. *Bioinformatics* **25**, 1105–1111 (2009).
34. Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotechnol.* **31**, 46–53 (2013).
35. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and cufflinks. *Nature Protocols* **7**, 562–578 (2012).
36. Xi, Y. & Li, W. BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinformatics* **10**, 232 (2009).
37. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18**, 1851–1858 (2008).
38. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357–359 (2012).
39. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).
40. Dai, Z. *et al.* shRNA-seq data analysis with edgeR. *F1000Res.* **3**, 95 (2014).
41. Smyth, G. K. in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor Statistics for Biology and Health* (ed. Gentleman, R.) (Springer, 2005).
42. Goff, L., Trapnell, C. & Kelley, D. cummeRbund: Analysis, Exploration, Manipulation, and Visualization of Cufflinks High-Throughput Sequencing Data. <http://compbio.mit.edu/cummeRbund/> (2012).
43. Li, Q. H., Brown, J. B., Huang, H. Y. & Bickel, P. J. Measuring reproducibility of high-throughput experiments. *Ann. App. Stat.* **5**, 1752–1779 (2011).
44. Zhang, Y. *et al.* Model-based analysis of ChIP-seq (MACS). *Genome Biol.* **9**, R137 (2008).
45. Mikkelsen, T. S. *et al.* Comparative epigenomic analysis of murine and human adipogenesis. *Cell* **143**, 156–169 (2010).
46. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).
47. Dabney, A. & Storey, J. D. *qvalue: Q-value estimation for false discovery rate control*. R package version 1.40.0 (2013).
48. Ross-Innes, C. S. *et al.* Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature* **481**, 389–393 (2012).
49. Neph, S. *et al.* An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489**, 83–90 (2012).
50. Zhu, L. J. *et al.* ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinformatics* **11**, 237 (2010).
51. Thompson, C. L. *et al.* A high-resolution spatiotemporal atlas of gene expression of the developing mouse brain. *Neuron* **83**, 309–323 (2014).
52. Fogel, G. B. *et al.* A statistical analysis of the TRANSFAC database. *Biosystems* **81**, 137–154 (2005).
53. Jolma, A. *et al.* DNA-binding specificities of human transcription factors. *Cell* **152**, 327–339 (2013).
54. Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L. & Noble, W. S. Quantifying similarity between motifs. *Genome Biol.* **8**, R24 (2007).
55. Bailey, T. L., Williams, N., Misleh, C. & Li, W. W. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.* **34**, W369–W373 (2006).
56. Manke, T., Roeder, H. G. & Vingron, M. Statistical modeling of transcription factor binding affinities predicts regulatory interactions. *PLOS Comput. Biol.* **4**, e1000039 (2008).
57. Manke, T., Heinig, M. & Vingron, M. Quantifying the effect of sequence variation on regulatory interactions. *Hum. Mutat.* **31**, 477–483 (2010).
58. Boulesteix, A. L. & Strimmer, K. Predicting transcription factor activities from combined analysis of microarray and ChIP data: a partial least squares approach. *Theor. Biol. Med. Model.* **2**, 23 (2005).
59. Boulesteix, A. L. & Strimmer, K. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Brief. Bioinform.* **8**, 32–44 (2007).
60. de Jong, S. Simpls: an alternative approach to partial least-squares regression. *Chemometr. Intell. Lab.* **18**, 251–263 (1993).
61. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
62. Sherwood, R. I. *et al.* Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nature Biotechnol.* **32**, 171–178 (2014).
63. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).

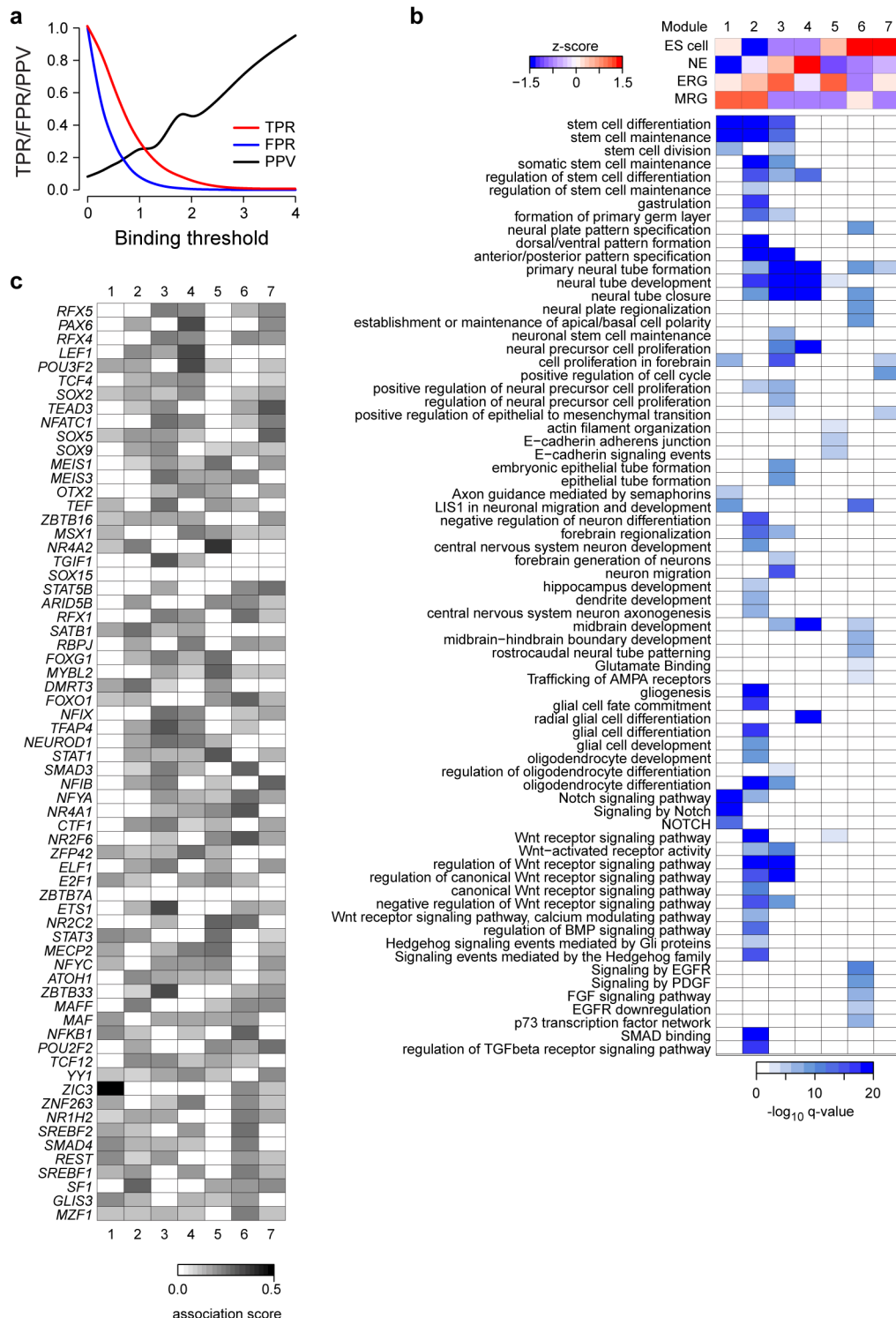


Extended Data Figure 1 | Isolation and characterization of ES-cell-derived neural progenitor cells. This figure relates to Fig. 1 in the main text.

a, Schematic of our differentiation model including the specific days of sample collection. Human ES cells were differentiated into NE cells using dual inhibition of TGF- β and bone morphogenetic protein followed by the transition to neural base media. Subsequently, sonic hedgehog and FGF8 are used to transition to the ERG stage. For the rest of the differentiation experiment the cells were constantly maintained in FGF2 and EGF2 neural base media to reach the MRG stage after 35 days (D35), the LRG stage after 80 and the LNP stage after about 200 days of *in vitro* culture. Cell type names indicated in red were profiled for gene expression, histone modifications as well as DName by WGBS, while names shown in grey for gene expression only and names in black for DName by RRBS only. NSB, noggin/SB-431542; SHH, sonic hedgehog; FGF, fibroblast growth factor; EGF, epidermal growth factor.

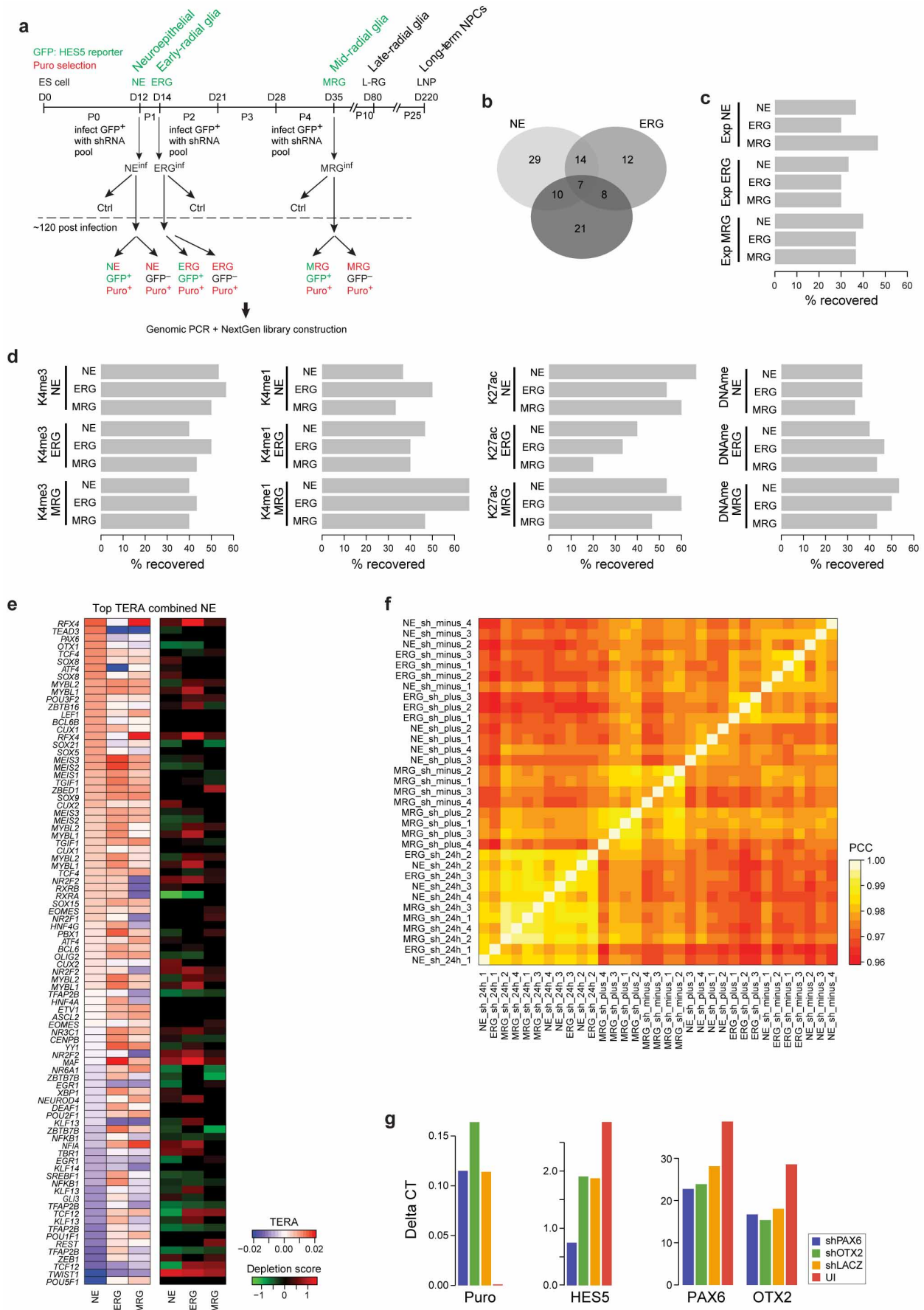
b, Hierarchical clustering for all RNA-seq data sets collapsing replicates using

the Jensen-Shannon (JS) divergence as a metric **c**, Gene expression patterns shown as z scores for all differentially expressed genes (q value ≤ 0.1) across ES cells and four neural precursor differentiation stages for genes expressed at ≥ 1 FPKM in at least one stage ($n = 20,306$). Genes were grouped into 18 clusters based on minimal average silhouette width using partitioning around medoids (PAM) clustering and Jensen-Shannon divergence as a metric. Pie charts below indicate the fraction of up- (red) and downregulated (green) genes during each transition. **d**, Gene expression patterns shown as z scores for all significantly differentially expressed genes (q value ≤ 0.1) across four more mature cell populations obtained through differentiation of NE, ERG or MRG cells to neuronal-like cells (NE/ERG/MRGdN) and astrocyte-like cells (LRGdA) derived from the LRG stage. Genes were grouped into 12 clusters based on minimal average silhouette width using PAM clustering and the Jensen-Shannon divergence as a metric.



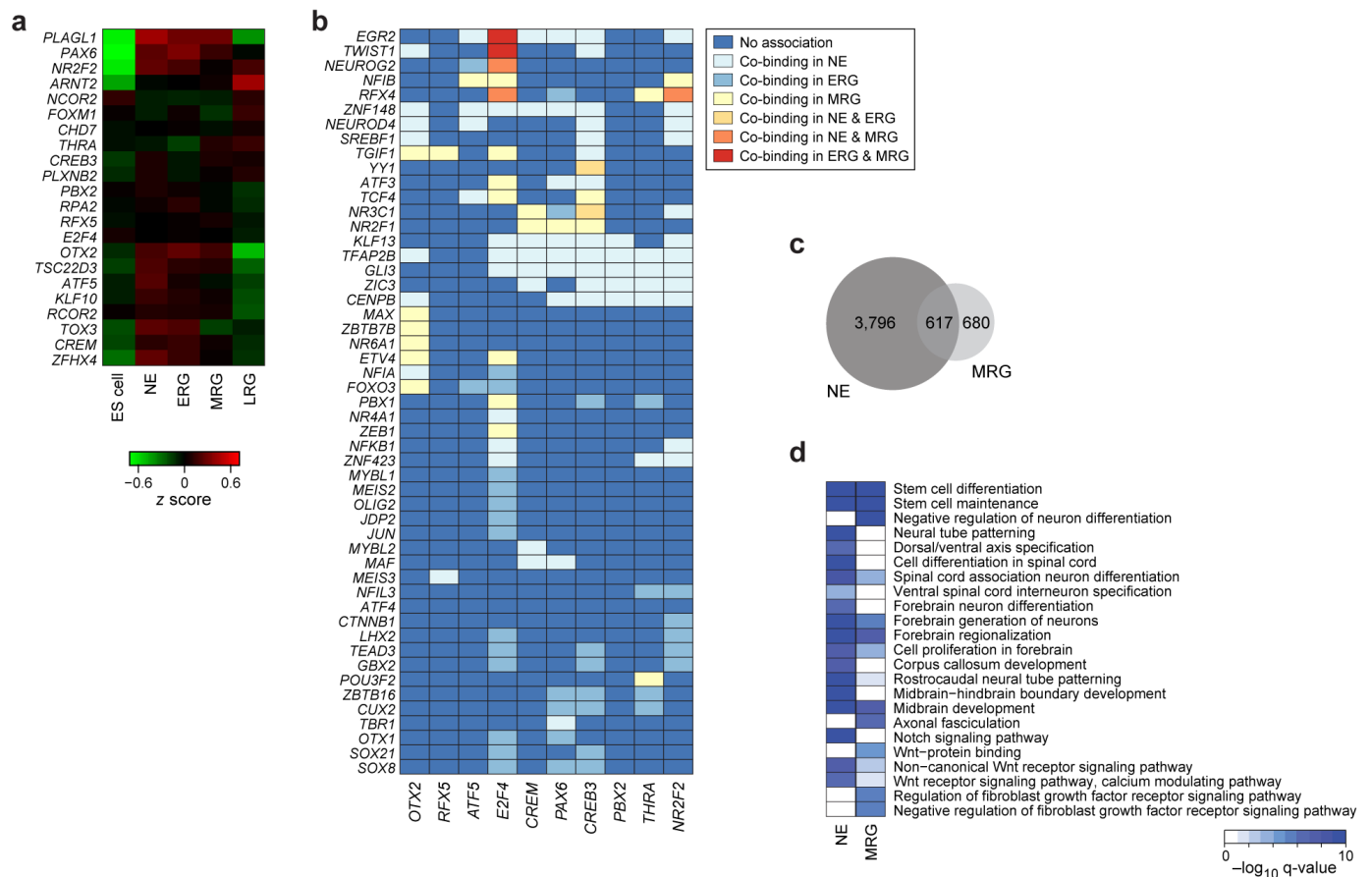
Extended Data Figure 2 | Epigenetic dynamics and transcription factor footprints. This figure relates to Fig. 2 in the main text. **a**, Median TPR (red), FPR (blue) and PPV (black) for $n = 46$ transcription factors with matching motif for H3K27ac footprints ($n = 27,292$) in K562 cells as a function of confidence in predicted binding ($-\log_{10} P$ value). True positives were defined as predicted binding events overlapping with peaks determined by ChIP-seq and false positives accordingly. The entire set of positives was defined as all transcription factor ChIP-seq peaks for a particular factor that overlapped with

any H3K27ac footprint. **b**, Top, decomposition of H3K27ac dynamics into 7 distinct modules based on PLS regression. Colours indicate median epigenetic enrichment level of gene regulatory elements assigned to each module for each cellular state for H3K27ac. Bottom, selected gene set enrichment analysis results for gene regulatory elements associated with each module. **c**, Connectivity matrix showing the association strength of each of the factors listed in Fig. 2b with each of the 7 modules identified by the partial least square (PLS) regression.



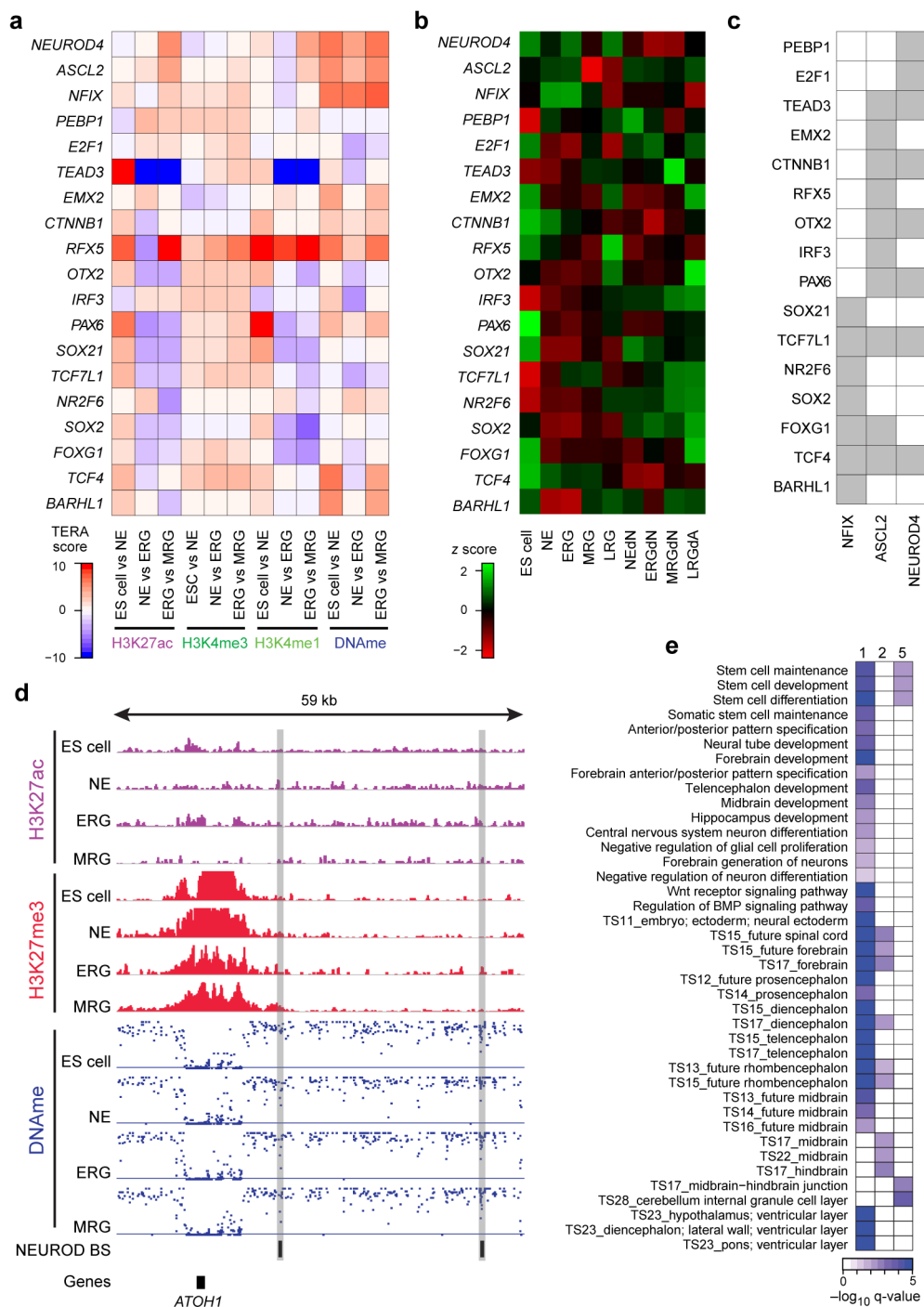
Extended Data Figure 3 | Functional validation using a pooled shRNA screen. This figure relates to Fig. 3 in the main text. **a**, Detailed outline of the pooled shRNA screen. Each stage (NE, ERG and MRG) was infected with an optimized virus titre aiming for an average of one shRNA integration per cell. Immediately after infection, cells were subjected to puromycin (puro) selection and bulk population material was collected 24 h after infection and before efficient shRNA knockdown. Five days after infection and selection, cells were FACS-sorted for HES5-GFP and both GFP⁺ and GFP⁻ cells were collected for analysis. Subsequently, gDNA was extracted and all integrated shRNAs were amplified by PCR for each population separately. The resulting material was then used to construct libraries for next-generation sequencing to count the number of shRNA integrations for each shRNA in each cell population. **b**, Overlap of genes identified to facilitate HES5⁺ cell maintenance, progression or proliferation determined by genes with at least two shRNAs significantly ($q \leq 0.05$) over-represented in the HES5⁺ population with respect to the 24-h or HES5⁻ control. **c**, Regulator predictions based on differential gene expression. Performance is measured as percentage of the top 20 differentially expressed factors for each stage for those the transcription factors

included in the shRNA library. **d**, Regulator predictions based on TERA ranking for H3K4me3, H3K4me1, H3K27ac or DNAm. Performance is measured as percentage of the top 20 predicted activating or repressive motifs for each stage mapping to a transcription factor included in the shRNA library. **e**, Detailed heat map showing the top 30 predicted motifs and corresponding transcription factors differentially active between consecutive differentiation stages based on the combined TERA scores for H3K27ac, H3K4me3, H3K4me1 and DNAm. In addition, knockdown results as depletion scores (green/red heat map) obtained at each stage are shown on the right. **f**, Heat map showing the pairwise Pearson correlation coefficient (PCC) of the log₂ read-count normalized shRNA libraries across all conditions and technical replicates. **g**, Individual validation for shRNAs against *OTX2* and *PAX6* at the NE stage, which showed no effect in our pooled screening approach at any stage. Shown are qPCR levels for *OTX2* or *PAX6*, *HES5* and puromycin relative to *HPRT*. Each gene was measured in an independent knockdown experiment for a pool of the five shRNAs against *PAX6* (blue), *OTX2* (green), *lacZ* (orange) as well as the uninfected control (red).



Extended Data Figure 4 | Co-binding analysis. This figure relates to Fig. 4 in the main text. **a**, Gene expression levels reported as z scores for core network transcription factors and epigenetic modifiers with and without a known DNA binding motif. **b**, Illustration of predicted significant co-binding relationships ($P \leq 0.01$, enrichment ≥ 1.5) of core factors (rows) with more stage-specific or pro-neuronal/glia factors (columns). Colour coding indicates

whether binding is stage specific or occurs at multiple stages. **c**, Overlap of predicted binding sites in dynamic putative enhancer regions based on H3K27ac for OTX2 in NE and ERG. **d**, Gene set enrichment analysis results for predicted OTX2 binding sites in dynamic putative enhancer regions at the NE and MRG stage.



Extended Data Figure 5 | Epigenetic priming. This figure relates to Fig. 5 in the main text. **a**, TERA scores for H3K27ac, H3K4me3, H3K4me1 and DNAm for transcription factors showing evidence of priming (top, bold) and transcription factors predicted to significantly co-occur in these primed binding sites. **b**, Gene expression levels shown as z scores for primed and co-binding transcription factors from panel **a**. **c**, Detailed predicted co-binding relationship ($P \leq 0.01$, enrichment ≥ 1.5) of primed transcription factors (columns) with significantly associated co-binding factors (rows).

d, Illustration of a potential priming event and the associated predicted target gene at the *ATOH1* locus (chromosome 4: 94,740–94,800). For each stage, H3K27ac, H3K27me3 and DNAm patterns are shown along with predicted NEUROD binding sites (black boxes) in putative gene regulatory elements marked by a loss of DNAm (highlighted by the grey bars). **e**, Gene set enrichment analysis results for predicted NEUROD binding sites split up by dynamic patterns defined in the top of Fig. 5b. Binding sites in patterns 3 and 4 showed no significant enrichment.

Cell-of-origin chromatin organization shapes the mutational landscape of cancer

Paz Polak^{1,2*}, Rosa Karlič^{3*}, Amnon Koren^{2,4}, Robert Thurman⁵, Richard Sandstrom⁵, Michael S. Lawrence², Alex Reynolds⁵, Eric Rynes⁵, Kristian Vlahoviček^{3,6}, John A. Stamatoyannopoulos^{5,7} & Shamil R. Sunyaev^{1,2}

Cancer is a disease potentiated by mutations in somatic cells. Cancer mutations are not distributed uniformly along the human genome. Instead, different human genomic regions vary by up to fivefold in the local density of cancer somatic mutations¹, posing a fundamental problem for statistical methods used in cancer genomics. Epigenomic organization has been proposed as a major determinant of the cancer mutational landscape^{1–5}. However, both somatic mutagenesis and epigenomic features are highly cell-type-specific^{6,7}. We investigated the distribution of mutations in multiple independent samples of diverse cancer types and compared them to cell-type-specific epigenomic features. Here we show that chromatin accessibility and modification, together with replication timing, explain up to 86% of the variance in mutation rates along cancer genomes. The best predictors of local somatic mutation density are epigenomic features derived from the most likely cell type of origin of the corresponding malignancy. Moreover, we find that cell-of-origin chromatin features are much stronger determinants of cancer mutation profiles than chromatin features of matched cancer cell lines. Furthermore, we show that the cell type of origin of a cancer can be accurately determined based on the distribution of mutations along its genome. Thus, the DNA sequence of a cancer genome encompasses a wealth of information about the identity and epigenomic features of its cell of origin.

Recent studies have begun to address the underlying causes of cancer mutational heterogeneity by comparing mutation rate variation to the distribution of sequence features, gene expression and epigenetic marks along the genome^{2–5}. A major limitation of previous studies was their uniform treatment of mutations from different cancers, and their consideration of epigenetic marks from a single cell type, usually a cell type different from the cancer tissue of origin. However, cancer is far from being a disease of uniform origin, progression and cell biology. Instead, different cancer types differ in their overall mutation rates, their predominant mutation types, and the distribution of mutations along their genomes¹. Substantial variation also exists in the epigenomic landscape of different tissues, specifically in patterns of chromatin accessibility, histone modifications, gene expression and DNA replication timing^{7–10}. Full understanding of the factors contributing to mutational heterogeneity in cancer genomes thus requires the evaluation of the relationship between multiple epigenetic marks and mutation patterns in a cell-type-specific manner.

We analysed a total of 173 cancer genomes from eight different cancer types that represent a wide range of tissues of origin, carcinogenic mechanisms, and mutational signatures: melanoma¹¹, multiple myeloma¹², lung adenocarcinoma¹³, liver cancer¹⁴, colorectal cancer¹⁵, glioblastoma¹⁶, oesophageal adenocarcinoma¹⁷ and lung squamous cell carcinoma¹⁸. Regional variations in mutation density appeared similar, although not identical, among the different cancer types (Extended Data Fig. 1).

We compared the genomic distribution of mutations in these cancer genomes to 424 epigenetic features that were measured by the Epigenome Roadmap consortium⁹. These features were derived from 106 different



cell types from 45 different tissue types, encompassing the established or likely cell types of origin of most of the cancer

types that we investigated (Methods and Extended Data Fig. 2). Notably, these data derive from primary human cells and tissues rather than malignant cell lines. These epigenetic features comprised eight different types of variables, including DNase I hypersensitivity (a global measure of chromatin accessibility)⁷ and various histone modifications. An example of the variation in mutation density along chromosomes at a 1 Mb scale together with the density of DNase I hypersensitive sites (DHSs) is shown in Fig. 1. In this case, as in most other cases (see later), epigenomics features indicative of active chromatin and transcription were associated with low mutation density, whereas repressive chromatin features were associated with regions of high mutation density. Notably, these statistical associations do not necessarily imply causal effects of individual chromatin features, nor point to specific biological mechanisms.

The comparison of individual epigenomic features with local mutation density revealed that the genomic distribution of chromatin features corresponding to the tumour's cell type of origin is more strongly associated with local mutation density than the distribution of features found in unrelated cell types. For example, DHSs from melanocytes explained a substantially larger fraction of the variance in melanoma mutation density than DHSs from other cell types, even from the same tissue (skin) (Fig. 1b). As another example, even though H3K4me1 marks in melanocytes and hepatocytes are highly correlated ($r = 0.8$), the distribution of mutations in liver cancer followed the levels of H3K4me1 in hepatocytes but not in melanocytes, whereas melanoma mutations correlated with the levels of H3K4me1 in melanocytes but not in hepatocytes (Fig. 1c).

This initial observation suggested that the impact of chromatin structure on local mutation density is highly cell-type specific. The comprehensive representation of different cell types in the Epigenome Roadmap could thus enable improved prediction accuracy of mutations compared to previous studies. To rigorously quantify the contribution of different chromatin marks and gene expression to regional mutation density, and the extent of cell type specificity, we used Random Forest regression (Methods).

Remarkably, epigenetic features, together with replication timing measured in ENCODE consortium cell lines¹⁹, collectively accounted for 74–86% of the variance in mutation density in seven cancer types (Fig. 2a). In glioblastoma, for which fewer mutations were available for the analysis, 55% of the variance in mutation density could be explained. This is substantially higher than in earlier studies⁴ and indicates that, at least for these cancer types, we have identified a set of epigenetic variables and cell types that almost fully predict the mutational variability along the genome. This enhanced prediction accuracy was not simply due to the larger size

¹Division of Genetics, Department of Medicine, Brigham & Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA. ²The Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA. ³Bioinformatics Group, Department of Molecular Biology, Division of Biology, Faculty of Science, University of Zagreb, Horvatovac 102a, 10000 Zagreb, Croatia. ⁴Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA. ⁵Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA. ⁶Department of Informatics, University of Oslo, P.O. Box 1080, Blindern, NO-0316 Oslo, Norway. ⁷Department of Medicine, Division of Oncology, University of Washington, Seattle, Washington 98195, USA.

*These authors contributed equally to this work.

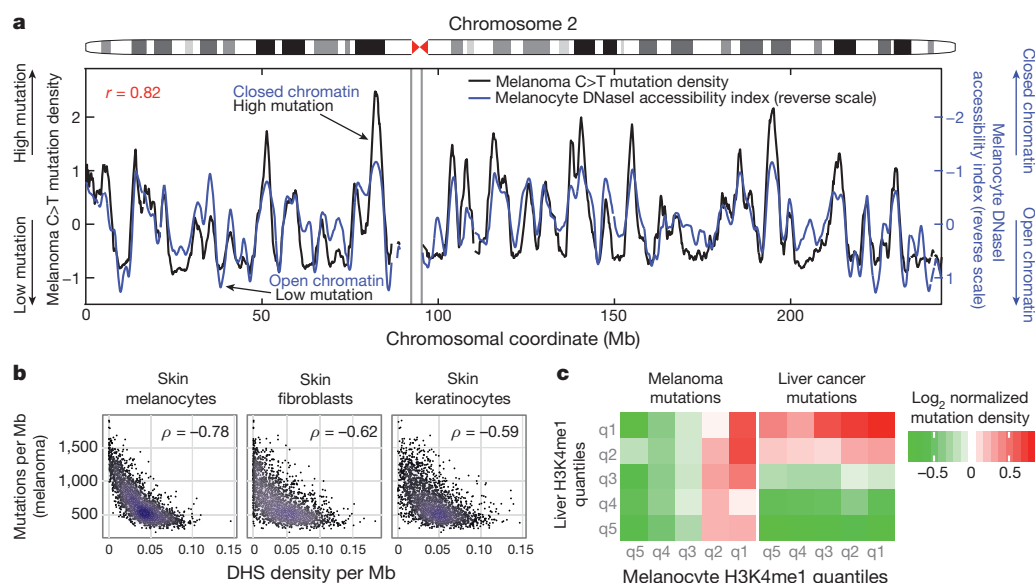


Figure 1 | Mutation density in melanoma is associated with individual chromatin features specific to melanocytes. **a**, The density of C > T mutations in melanoma alongside a 100-kb window profile of melanocyte chromatin accessibility ('DNase I accessibility index'; shown in normalized, reverse scale; high values correspond to less accessible chromatin and vice versa). **b**, The number of mutations per megabase in melanoma versus DHS

of the training data relative to previous studies, as the predictive ability dropped by only ~2–6% when only 10% of the data was used (Extended Data Fig. 3).

Prediction accuracy in individual samples is expected to be lower than in samples pooled by cancer type due to tumour heterogeneity, sampling variance, and a lower number of mutations available for the analysis (Extended Data Fig. 4). To evaluate the influences of these variables on the prediction ability of the Random Forest model, we simulated mutation data sets of variable sizes generated by the model itself, and compared the prediction accuracy of simulated and real data as a function of the number of mutations. For most samples, epigenomic features explained most (on average 70%) of the maximally predicted variance (Extended

density, for three types of skin cells. The Spearman's rank correlations (ρ) are shown on each panel. **c**, The normalized density of mutations in liver cancer and melanoma genomes as a function of density quintiles of H3K4me1 marks in liver cells and in melanocytes. For both cancer genomes, mutation density depends only on H3K4me1 marks measured in the cell of origin.

Data Fig. 5), and more than was explained by earlier studies^{2,4} when matching data set sizes. As a point of direct comparison with an earlier study² that did not use cell-type-specific chromatin marks, our model explained 50% of the variance in mutation density in the melanoma cell line COLO829 (ref. 20), compared with 29% in the earlier study.

The prediction accuracy was similar whether testing for all mutations or only the mutations of the predominant type (Extended Data Fig. 6) in each cancer type^{1,21} (Fig. 2b). A notable exception was lung adenocarcinoma, where a larger fraction of the variance could be explained for G > T mutations associated with smoking^{13,22} than for C > T mutations. This difference was observed for both samples with G > T transversions¹³ and C > T transitions as the leading mutational sources (Fig. 2c).

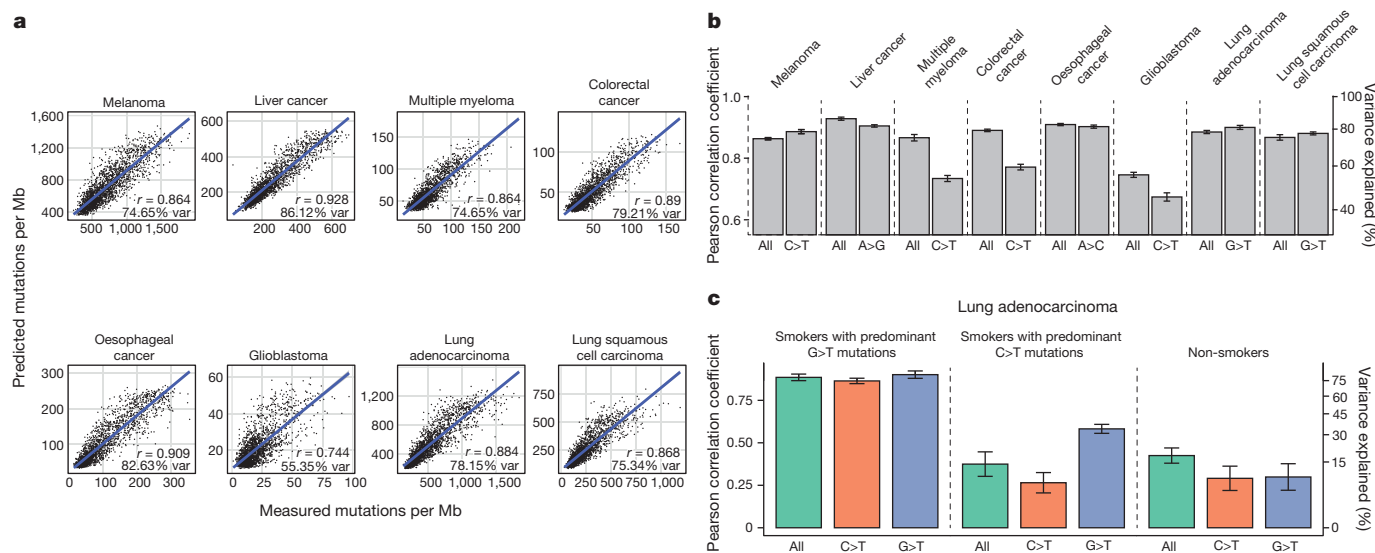


Figure 2 | Predicting local mutation density in cancer genomes using Random Forest regression trained on 424 epigenomic profiles. Pearson correlation between observed and predicted mutation densities along chromosomes is shown. **a**, Actual versus predicted mutation densities in eight cancers. **b**, **c**, Prediction accuracy represented as mean \pm s.e.m.

(estimated using tenfold cross-validation). Panels show prediction accuracy for all mutations and for nucleotide changes predominant in the corresponding cancer (**b**), and prediction accuracy in lung adenocarcinoma genomes stratified by smoking history and predominant nucleotide changes (G>T or C>T) (**c**).

Prediction accuracy was fully explained by chromatin features, with gene expression and nucleotide content not providing any further improvement to the accuracy of the model. Even though gene expression has been unequivocally demonstrated to influence mutation density, chromatin features appear to be statistically stronger predictors (Extended Data Fig. 7).

When considering individual feature contributions to mutation rate prediction, between six and nineteen variables passed the significance threshold in any individual cancer type. There was a sweeping association between cancer mutations and chromatin marks measured in the

cell type of origin of each cancer (Fig. 3a). For instance, six out of the top ten features explaining variation in melanoma mutation density were derived from melanocytes (Figs 1 and 3b). Similarly, seven out of the nine top features explaining mutational profiles in liver cancer were measured in liver cells. Comparable results were obtained for multiple myeloma, colorectal adenocarcinoma and glioblastoma, where most of the significant features were measured in haematopoietic, intestine mucosa and brain tissues, respectively. For oesophageal adenocarcinoma, the top predictors were chromatin features derived from stomach mucosa rather than from oesophageal tissues; this is expected given

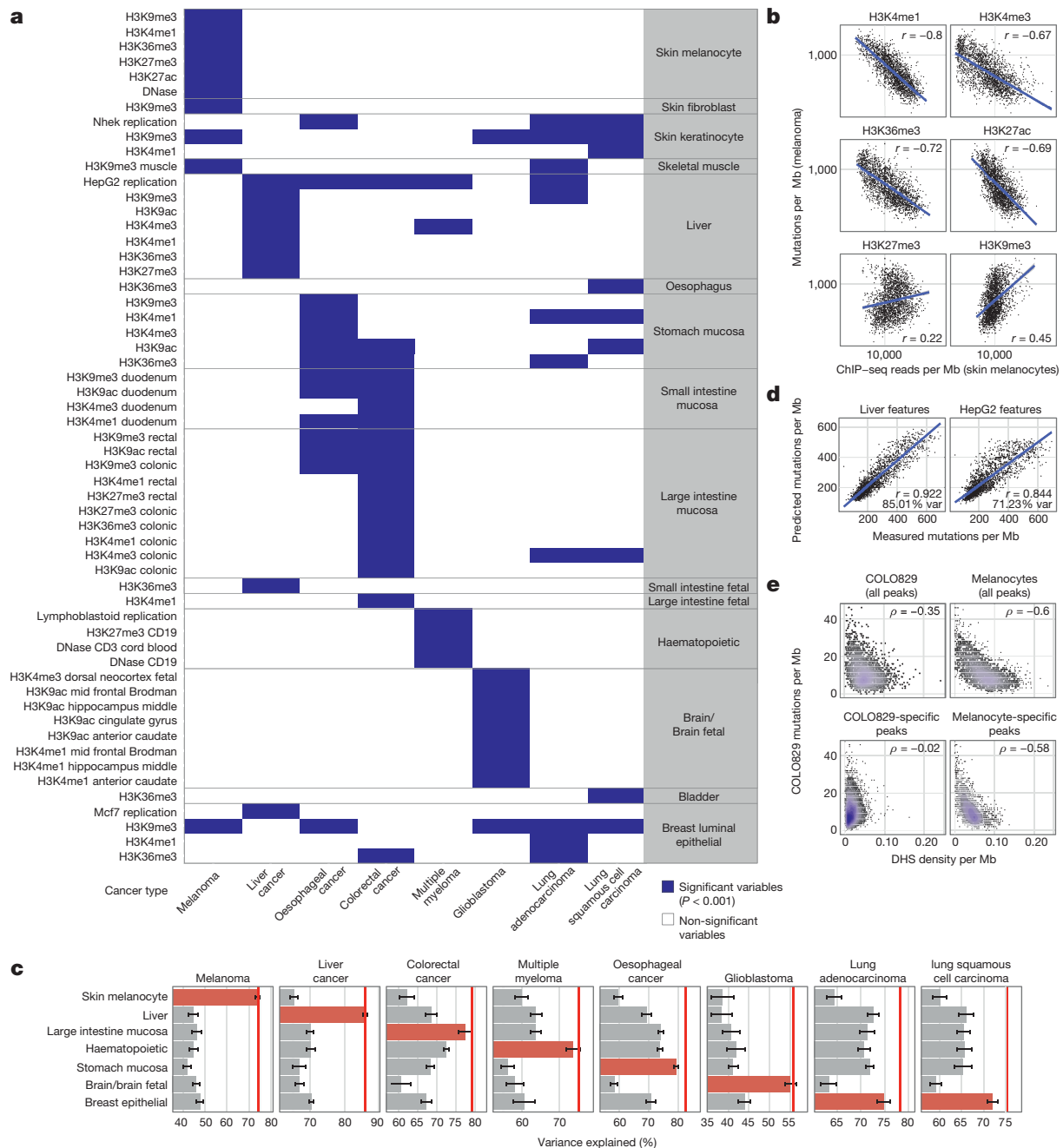


Figure 3 | Epigenomic features that significantly contribute to the prediction of local mutation density. **a**, Features (blue rectangles) that significantly contributed to the predictions in at least one cancer type (see Methods). **b**, Melanoma mutation density versus the density of chromatin modifications in melanocytes. **c**, Prediction accuracy (mean \pm s.e.m. estimated using tenfold cross-validation) of models separately trained on features from different tissues for each cancer type. Red bars, tissues with the highest

prediction accuracy. Red line, prediction accuracy when using all 424 epigenetic features. **d**, Comparison of predictions accuracies of liver cancer mutation density from features of normal liver cells versus cancer cells (HepG2). **e**, Mutation density in COLO829 melanoma cell line versus DHS density in COLO829, melanocytes, DHSs specific to COLO829 (not observed in melanocytes) and DHSs specific to melanocytes (not observed in COLO829). Spearman's rank correlation coefficient is given for each comparison.

that the analysed oesophageal adenocarcinomas were triggered by Barrett's oesophagus cells that resemble stomach epithelial cells²³. Lung adenocarcinoma and lung squamous cell carcinoma were the only exceptions in that the top predictors were scattered among different tissue groups; the lack of tissue specificity in these cases likely results from the absence of epigenomic marks from normal lung epithelial cells in our data set.

The results of the Random Forest regression were confirmed using backward feature selection to identify the minimal set of epigenetic predictors of mutations in each cancer type (Methods). As few as three to five features were sufficient to capture the variance explained by the full set of 424 different features (Extended Data Fig. 8), and in all cancers besides lung (as above), most of these features were derived from the corresponding cell types of origin. As a more direct test, we grouped all epigenomic data by cell or tissue type and compared the collective explanatory power of chromatin features derived from the cell types of origin versus unmatched cell types. The results of this analysis confirmed the cell type specificity of the association between chromatin features and mutation density (Fig. 3c).

The above results pose a key question on whether epigenomic features derived from the cell type of origin are the strongest determinants of cancer mutations or whether they simply serve as the best available proxies to the chromatin organization of the corresponding malignant cells. The availability of epigenomic data for the liver cancer cell line HepG2 (ref. 8) and for melanoma cell lines made it possible to directly address this question. Surprisingly, in both cases, epigenomic features from the cell type of origin resulted in a higher prediction accuracy than those from the cancer cell lines. The Random Forest predictor trained on chromatin features of HepG2 was less accurate in predicting the liver cancer mutation density than the analogous predictor trained on features of hepatocytes (Fig. 3d). Similarly, chromatin accessibility in melanocytes was a much better predictor of mutation density in the COLO829 melanoma cell line (Fig. 3e and Extended Data Fig. 9). Thus, chromatin features associated with carcinogenesis do not determine cancer mutations to the same extent as chromatin features of the cells of origin. We envision two potential explanations for this observation. First, most of the somatic mutations

observed in cancers may arise before the epigenetic changes linked to neoplastic progression. Second, advanced tumours may undergo specific epigenetic changes that distinguish them from other tumours of the same type.

Taken together, the above results strongly suggest that the cell of origin of an individual tumour sample could be predicted from its mutation pattern alone. Mutation profiles of individual samples cluster according to cancer type, and, consequently cell of origin (Fig. 4a). We developed a straightforward predictor based on enrichment of epigenomic variables from a single cell type among the top 20 variables selected by the Random Forest analysis. This approach classified 88% of melanoma, colorectal, liver, multiple myeloma, oesophageal and glioblastoma cancer genomes to melanocytes, colonic mucosa, liver, haematopoietic, stomach mucosa and brain tissues, respectively (Fig. 4b). Thus, mutational patterns contain sufficient information for identifying the cell type of origin of a tumour. We propose that sequencing the DNA of a tumour of unknown primary origin can allow the precise identification or categorization of the cell type of origin of that tumour.

Traditionally, statistical prediction in cancer has made use of gene expression data. We therefore constructed an analogous predictor of cell of origin using RNA sequencing data from 167 glioblastoma multiforme and 370 skin cutaneous melanoma samples²⁴. This predictor achieved accuracies of 78% and 57% on these cancer types, slightly lower than the mutation-based predictor. Although these two classifiers are not directly comparable, it is clear that genome sequence carries at least the same amount of information about the cell of origin as gene expression data does.

In conclusion, our observations suggest that cancer mutation density is linked to the epigenomic profile in a highly cell-type-specific manner. Thus, DNA sequence is informative about the origin of an individual tumour. The accumulating epigenomic data on human cell types opens the perspective for accurate prediction of the cell of origin of a cancer from its genome sequence.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 6 December 2013; accepted 7 January 2015.

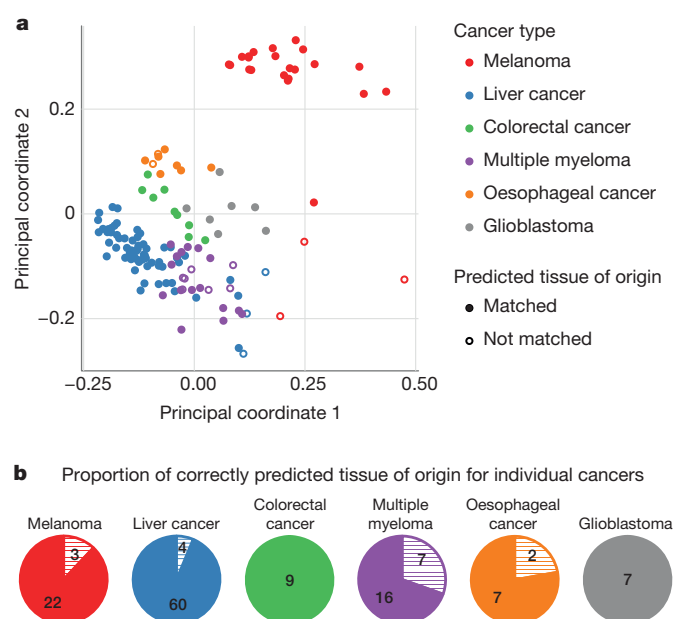


Figure 4 | Analysis of individual cancer genomes and prediction of cell type of origin. **a**, Principal coordinate analysis (PCOA) of the distribution of mutations in individual cancer genomes. Filled circles represent cancers for which the correct cell type of origin was identified. **b**, The accuracy of cell type of origin prediction for individual cancer genomes: the number of cancer samples that were assigned to the correct (solid colours) or incorrect (textured) cell types of origin based on their mutation profile.

- Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
- Hodgkinson, A., Chen, Y. & Eyre-Walker, A. The large-scale distribution of somatic mutations in cancer genomes. *Hum. Mutat.* **33**, 136–143 (2012).
- Liu, L., De, S. & Michor, F. DNA replication timing and higher-order nuclear organization determine single-nucleotide substitution patterns in cancer genomes. *Nature Commun.* **4**, 1502 (2013).
- Schuster-Böckler, B. & Lehner, B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* **488**, 504–507 (2012).
- Woo, Y. H. & Li, W. H. DNA replication timing and selection shape the landscape of nucleotide variation in cancer genomes. *Nature Commun.* **3**, 1004 (2012).
- Zhu, J. *et al.* Genome-wide chromatin state transitions associated with developmental and environmental cues. *Cell* **152**, 642–654 (2013).
- Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).
- The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
- Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* <http://dx.doi.org/10.1038/nature14248> (this issue).
- Ryba, T. *et al.* Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Res.* **20**, 761–770 (2010).
- Berger, M. F. *et al.* Melanoma genome sequencing reveals frequent *PREX2* mutations. *Nature* **485**, 502–506 (2012).
- Chapman, M. A. *et al.* Initial genome sequencing and analysis of multiple myeloma. *Nature* **471**, 467–472 (2011).
- Imielinski, M. *et al.* Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* **150**, 1107–1120 (2012).
- Totoki, Y. *et al.* High-resolution characterization of a hepatocellular carcinoma genome. *Nature Genet.* **43**, 464–469 (2011).
- Bass, A. J. *et al.* Genomic sequencing of colorectal adenocarcinomas identifies a recurrent *VTI1A-TCF7L2* fusion. *Nature Genet.* **43**, 964–968 [10.1038/ng.936](http://dx.doi.org/10.1038/ng.936) (2011).
- Brennan, C. W. *et al.* The somatic genomic landscape of glioblastoma. *Cell* **155**, 462–477 (2013).
- Dulak, A. M. *et al.* Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. *Nature Genet.* **45**, 478–486 (2013).

18. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519–525 (2012).
19. Hansen, R. S. *et al.* Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc. Natl Acad. Sci. USA* **107**, 139–144 (2010).
20. Pleasance, E. D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191–196 (2010).
21. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
22. Pleasance, E. D. *et al.* A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463**, 184–190 (2010).
23. Reid, B. J., Li, X., Galipeau, P. C. & Vaughan, T. L. Barrett's oesophagus and oesophageal adenocarcinoma: time for a new synthesis. *Nature Rev. Cancer* **10**, 87–101 (2010).
24. Hudson, T. J. *et al.* International network of cancer genome projects. *Nature* **464**, 993–998 (2010).

Acknowledgements This work was supported by NIH grants R01 MH101244, U54 CA143874 to S.R.S. and U01 ES017156, P01 HL53750, U54 HG007010 to J.A.S. R.K. and K.V. acknowledge the Integra-Life Seventh Framework Program (grant number

315997) and the EMBO Young Investigator Program (Installation grant 1431/2006 to K.V.).

Author Contributions S.R.S., J.A.S., P.P. and R.K. conceived the project and provided leadership. P.P., R.K., A.K., M.S.L., R.S., A.K., R.T., E.R., A.R. and K.V. analyzed the data and contributed to scientific discussions. S.R.S., P.P., R.K., A.K. and J.A.S. wrote the paper.

Author Information Epigenomic data are available from the NCBI via the GEO series GSE18927 for University of Washington Human Reference Epigenome Mapping Project. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.R.S. (ssunyaev@rics.bwh.harvard.edu) or J.A.S. (jstam@u.washington.edu).



This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported licence. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons licence, users will need to obtain permission from the licence holder to reproduce the material. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-sa/3.0>

METHODS

Data. We divided the human genome into 1 Mb regions, excluding regions overlapping centromeres and telomeres, as well as regions where the fraction of uniquely mappable base pairs was lower than 0.92. We calculated the mean signal for different histone modifications, DNase I hypersensitivity and replication timing in different cell types, and used these 424 features to predict mutation density along the genome in eight different cancer types (see below).

We calculated mutation density by obtaining data for 173 individual cancer genomes, belonging to eight cancer types: melanoma (25 genomes)¹¹, lung adenocarcinoma (24 genomes)¹³, lung squamous cell carcinoma (12 genomes)¹⁸, oesophageal adenocarcinoma (9 genomes)¹⁷, liver (64 genomes)¹⁴, multiple myeloma (23 genomes)¹², colorectal cancer¹⁵ (CRC, 9 genomes) and glioblastoma (7 genomes)¹⁶. The whole genome of the COLO-829 cell line has been sequenced by the Sanger Institute. The COLO-829 cell line was derived from metastatic tissue. The liver cancers were sequenced by the National Cancer Center Research Institute in Japan. The mutation lists for the COLO829 cell line and liver cancer that we used in this study can be found at (http://dcc.icgc.org/repository/legacy_data_releases/version_07/) under the folders Malignant_Melanoma-WTSI-UK (COLO-829) and Liver_Cancer-NCC-JP. The rest of the genomes were sequenced and analysed by the Broad Institute and called using MuTect²⁵ (<http://www.broadinstitute.org/cancer/cga/mutect>).

For each cancer type we counted the overall number of mutations in all individual cancer genomes belonging to that cancer type. We also determined the mutation densities for all possible types of mutations in each cancer types by counting different types of mutations in 1-Mb windows and normalizing for the sequence composition of each window.

We downloaded data for 7 different histone modifications and DNase I hypersensitivity from Epigenomics Roadmap⁹ and ENCODE⁸ (Extended Data Fig. 1). Epigenomic data are available from the NCBI via the GEO series GSE18927 for University of Washington Human Reference Epigenome Mapping Project at (<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE18927>). Data used in this study can also be viewed via multiple browsers outlined at the (<http://roadmapepigenomics.org/>) website.

Fetal tissues were obtained from morphologically normal fetuses by the Birth Defects Research Laboratory in the Department of Paediatrics at the University of Washington, collected under an IRB-approved protocol. Blood cell subsets were collected from fully consenting, normal donors at the Cellular Therapy Laboratory and cGMP Cell Processing Facility under the direction of Shelly Heimfeld at the Fred Hutchinson Cancer Research Center with IRB-approval.

For histone modifications we combined reads for all samples belonging to one cell type and calculated RPKM values for 1-Mb windows along the genome. We also calculated the average number of DNase I hypersensitivity peaks overlapping 1-Mb windows across all samples belonging to a certain cell type. We used BEDOPS²⁶ to map reads and DHS peaks to intervals.

We obtained data for four different Repli-seq experiments from the ENCODE project (Extended Data Fig. 1) and determined replication timing as the average value of wavelet-smoothed signal in each 1-Mb window. Lymphoblastoid cell line replication time was obtained from ref. 27 and averaged over 1-Mb windows along the genome.

To control for the effect of sequence features on mutation density, for each 1-Mb window we also calculated GC content, the number of CpG, GpC, and ApT dinucleotides, and fraction of the window overlapping coding regions, known genes and CpG islands.

To control for the effect of expression on mutation density we downloaded mRNA-seq data from the Epigenomics Roadmap⁹, for 38 different cell types for which expression data was available (Extended Data Table 1). We combined reads for all samples belonging to one cell type and calculated reads per kilobase per million mapped reads (RPKM) values for the set of all protein coding exons in 1-Mb windows, the set of all protein coding and lncRNA exons in 1-Mb windows, the maximally expressed gene in a 1-Mb window or non-genic regions in 1-Mb windows.

Random Forest regression. Random Forest is a non-parametric machine learning method that combines the output of an ensemble of regression trees to predict the value of a continuous response variable²⁸. The use of multiple regression trees reduces the risk of over-fitting and makes the method robust to outliers and noise in the input data. For each regression tree, a training set of n observations are drawn, with replacement, from the data set. The remaining data (out-of-bag data) constitutes the test set for this tree, and is used to compute the mean squared prediction error of the tree. The prediction for each observation is made by taking the average of predictions over all trees for which the observation was part of the out-of-bag data.

Random Forest provides an internal measure of the importance of different predictor variables, based on out-of-bag data. The mean squared error calculated on the out-of-bag data is recorded in every tree grown in the forest. The values of all the predictor variables are then randomly permuted in all the out-of-bag observations and the mean squared error is computed again. The difference between the

two errors is averaged over all the trees, and normalized by the standard error, representing the raw importance score for each variable.

We used Random Forest with 1,000 trees to predict mutation densities in 1Mb non-overlapping windows in the eight different cancer types using 424 predictor variables (epigenetic features and replication timing; Extended Data Fig. 1). We divided the data into ten non-overlapping sets and predicted the number of mutations in each cancer type using tenfold cross-validation. For each sample, the predicted value corresponded to the predicted mutation density when this observation was part of the test set. We used Pearson product-moment correlation to interpret the prediction accuracy. The fraction of variance explained by each model was calculated as the Pearson correlation coefficient squared.

Controlling for the effect of sequence features and expression on prediction accuracy. We created different subsets of features corresponding to chromatin (histone modifications and DNase hypersensitivity, 419 features), replication timing (5 features), sequence (7 features) and expression (38 features). We then used Random Forest regression with tenfold cross-validation to predict mutation density in different cancers, where for each cancer type we trained different models: on each subset of features separately and on combinations of different subsets of features.

Variable importance analysis. Variable importance was calculated for each predictor variable in each cancer type by permuting the variable, that is, randomly shuffling the data values so that the relationship between the response and predictor variables was destroyed. The percent of increase in mean squared error of prediction was then calculated. Since the variable importance can be influenced by both the correlation and the scale of the variables, we calculated the empirical P value of variable importance measures by repeatedly permuting the response variable in Random Forest models, in order to determine the distribution of measured importance values for each predictor variable²⁹. This procedure was repeated 1,000 times, and the number of times in which the importance measure in the original data set was lower or equal to the permuted importance measure was counted; this count represented the P value, with a count of one corresponding to a significance level of $P < 0.001$.

Feature selection. We applied backward elimination to identify a minimal set of predictors for each cancer type. Backward elimination is a 'greedy' algorithm which finds the locally optimal subset of features, but does not guarantee finding the global optimum. However, it is less computationally intensive than searching all possible feature subsets when the number of features (p) is large (in our case $p = 424$). Initially, we trained a Random Forest with tenfold cross-validation on the complete set of variables and determined the importance of all the variables in the model (the importance was calculated as the mean importance of the variable across 10 rounds of cross-validation). We then ranked the variables according to their importance and determined the top 20 variables. We then sequentially trained 20 models, removing the least important variable at each step, until only one predictor variable was left for training.

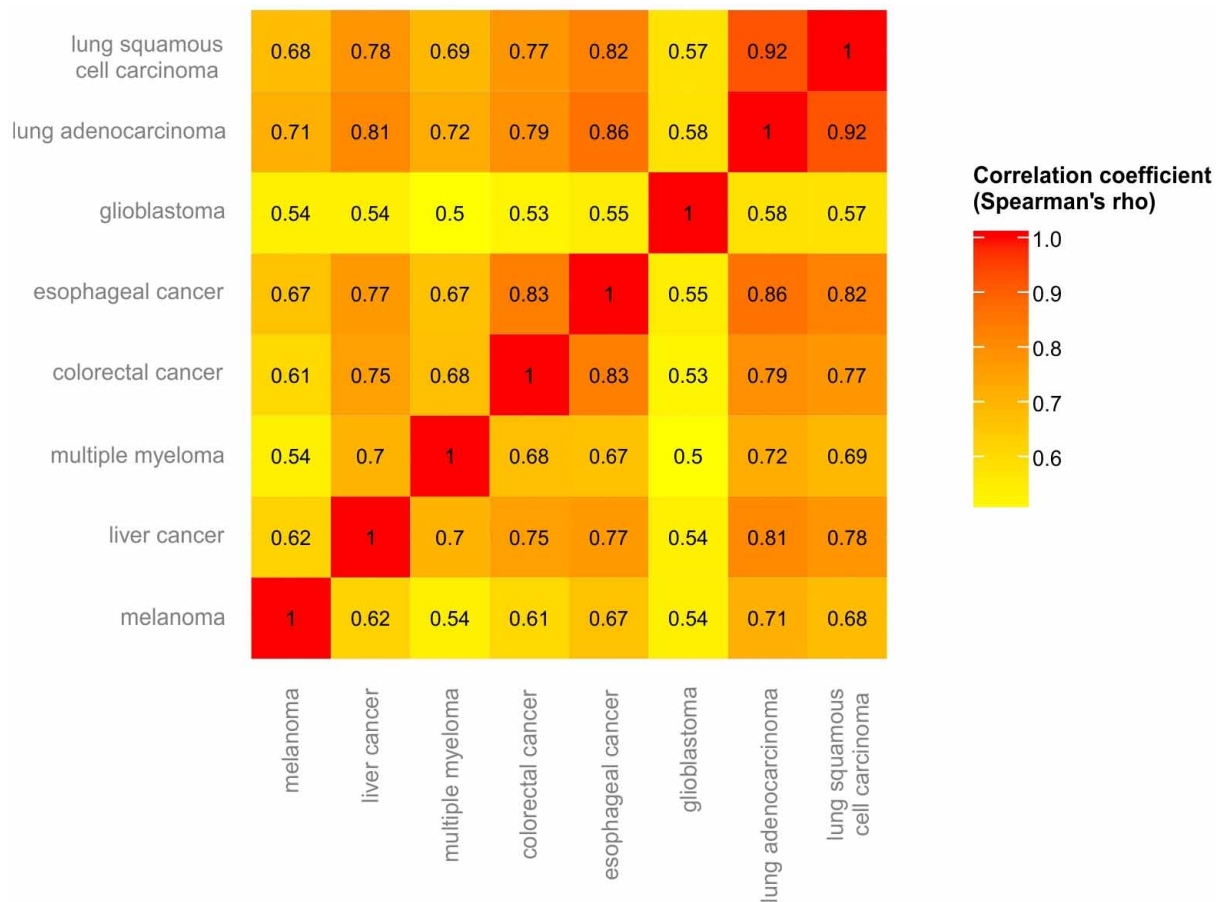
Principal coordinate analysis. We used principal coordinate analysis to visualize the dissimilarities in mutation density distributions between individual cancer genomes. Dissimilarity was calculated as $1 - \text{Pearson correlation coefficient}$, for all possible combinations of individual cancer genomes.

Prediction of tissue of origin for individual cancer genomes. For each individual cancer genome we predicted the density of mutations using Random Forest regression with tenfold cross-validation. We used the full set of features and determined the top 20 features according to the variable importance measure. We then calculated the enrichment of each tissue type among the top 20 features using the hypergeometric test and chose the tissue showing the most significant enrichment as the most likely tissue of origin for the individual cancer genome. We then calculated the percentage of individual cancers where the assigned tissue of origin matched the predicted tissue of origin.

Prediction of cell of origin using gene expression. For each individual cancer we downloaded gene expression data from The Cancer Genome Atlas¹⁶ and calculated the expression of the same genes in the 38 cell types for which mRNA-seq data was available from the Epigenomics Roadmap (Extended Data Table 1). For each cancer we trained a Random Forest regression model in which the gene expression values in cancer were used as the response variable and the gene expression in normal cells as the predictors. We identified the predictor variable, which showed the highest value of variable importance in the model and assigned the corresponding cell type as cell of origin of the cancer.

Sample size. No statistical methods were used to predetermine sample size.

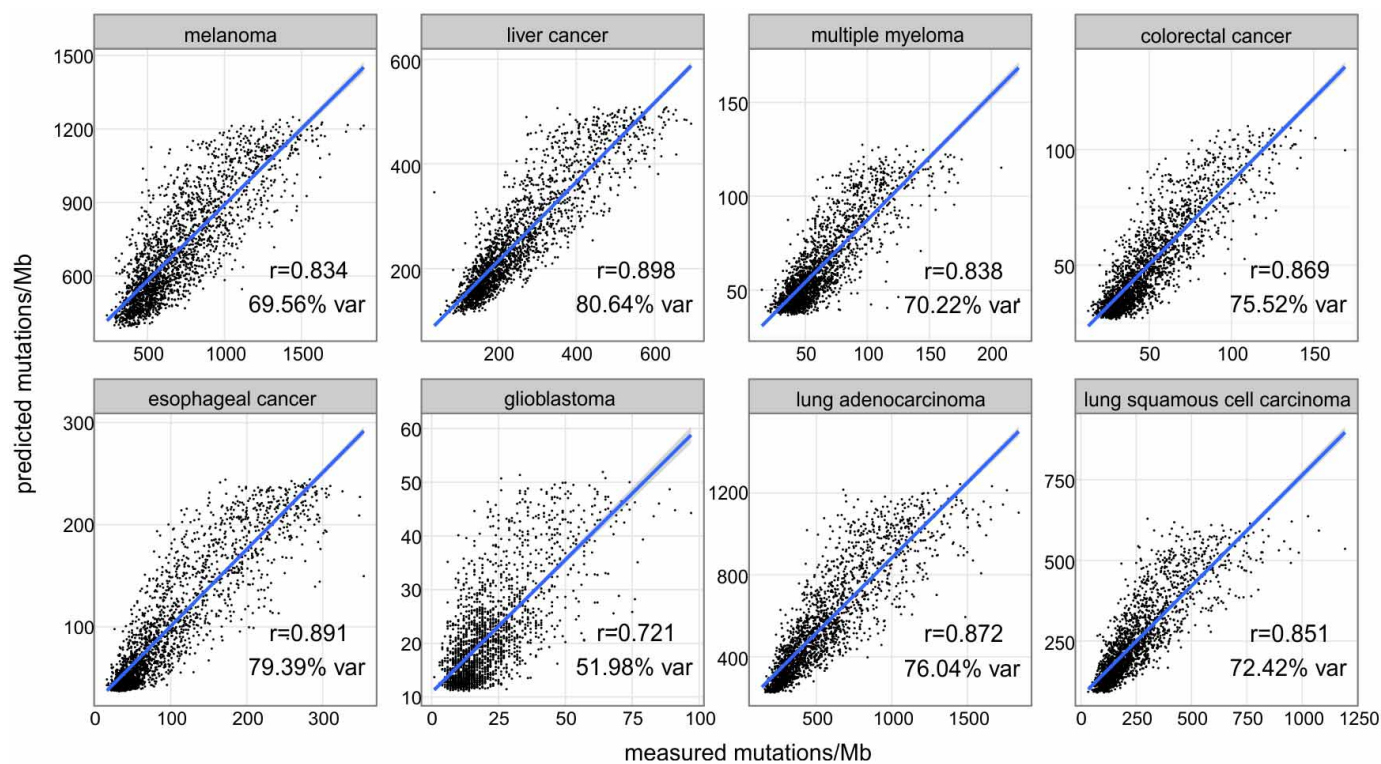
25. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnol.* **31**, 213–219 (2013).
26. Neph, S. *et al.* BEDOPS: high-performance genomic feature operations. *Bioinformatics* **28**, 1919–1920 (2012).
27. Koren, A. *et al.* Differential Relationship of DNA replication timing to different forms of human mutation and variation. *Am. J. Hum. Genet.* **91**, 1033–1040 (2012).
28. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).
29. Altmann, A., Tolosi, L., Sander, O. & Lengauer, T. Permutation importance: a corrected feature importance measure. *Bioinformatics* **26**, 1340–1347 (2010).



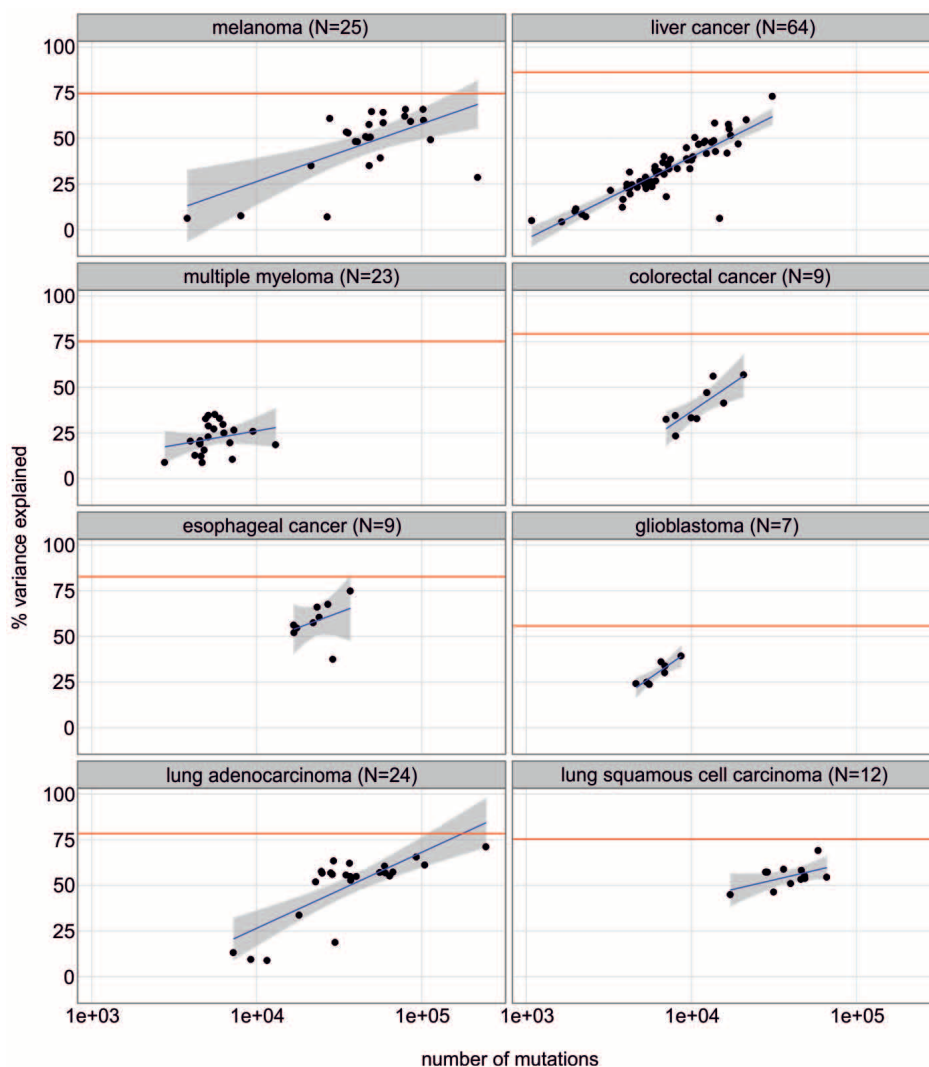
Extended Data Figure 1 | Correlation of mutation density measured in different cancer types.

Cell type	DNase-seq	H3K27ac	H3K27me3	H3K36me3	H3K4me1	H3K4me3	H3K9ac	H3K9me3	replication timing		
adrenal gland										adrenal gland	
adrenal gland fetal										adrenal gland fetal	
bladder										bladder	
brain mid frontal Brodmann area 9/46 dorsolateral prefrontal cortex										brain	
brain hippocampus middle											
brain cingulate gyrus											
brain anterior caudate											
brain angular gyrus											
brain germinal matrix fetal										brain fetal	
brain fetal											
brain dorsal neocortex fetal										breast	
Mcf7											
breast vHMEC										breast epithelial	
breast myoepithelial cells											
breast luminal epithelial cells										esophagus	
esophagus											
H1 derived neuronal progenitor cultured cells										H1	
H1 derived mesenchymal stem cells											
H1 cell line											
H1 BMP4 derived trophoblast cultured cells										H9 cell line	
H1 BMP4 derived mesendoderm cultured cells											
H9 cell line										heart	
heart										heart fetal	
heart fetal										hematopoietic	
lymphocyte											
CD8 primary cells											
CD8 mobilized primary cells											
CD56 primary cells											
CD56 mobilized primary cells											
CD4 primary cells											
CD4 mobilized primary cells											
CD34 primary cells											
CD34 mobilized primary cells											
CD3 primary cells											
CD3 mobilized primary cells											
CD3 cord blood primary cells											
CD20 primary cells											
CD19 primary cells											
CD14 primary cells											
IMR90											
iPS DF 6											
iPS DF 4											
iPS DF 19											
kidney										IMR90 cell line	
kidney right fetal										iPS	
kidney renal pelvis right fetal											
kidney renal pelvis left fetal										kidney fetal	
kidney renal pelvis fetal											
kidney renal cortex right fetal											
kidney renal cortex left fetal											
kidney renal cortex fetal											
kidney left fetal											
kidney fetal											
sigmoid colon											large intestine
colon smooth muscle											large intestine fetal
large intestine fetal											large intestine mucosa
rectal mucosa										liver	
colonic mucosa										lung	
liver										lung fetal	
HepG2											
lung										muscle fetal	
lung right fetal											
lung left fetal											
lung fetal											
muscle upper trunk fetal											
muscle upper limb fetal										ovary	
muscle upper back fetal											
muscle trunk fetal											
muscle lower limb fetal											
muscle leg fetal											
muscle back fetal											
muscle arm fetal											
ovary											
ovary fetal										ovary fetal	
pancreas										pancreas	
placenta day91										placenta	
placenta day85											
placenta day113											
placenta day108											
placenta day105											
psoas muscle										psoas muscle	
skeletal muscle										skeletal muscle	
skin upper back fetal										skin fetal	
skin scalp fetal											
skin quadriceps right fetal											
skin quadriceps left fetal											
skin fetal											
skin biceps right fetal											
skin biceps left fetal											
skin back fetal											
skin abdomen fetal											
penis foreskin fibroblast primary cells											
penis foreskin keratinocyte primary cells										skin fibroblast	
Nhek										skin keratinocyte	
penis foreskin melanocyte primary cells										skin melanocyte	
small intestine										small intestine	
small intestine fetal										small intestine fetal	
duodenum mucosa										sm. intestine mucosa	
spinal cord fetal										spinal cord fetal	
spleen fetal										spleen fetal	
gastric										stomach	
stomach fetal										stomach fetal	
stomach mucosa										stomach mucosa	
testes fetal										testes fetal	
thymus										thymus	
thymus fetal										thymus fetal	

Extended Data Figure 2 | Chromatin features and replication data used in the models.

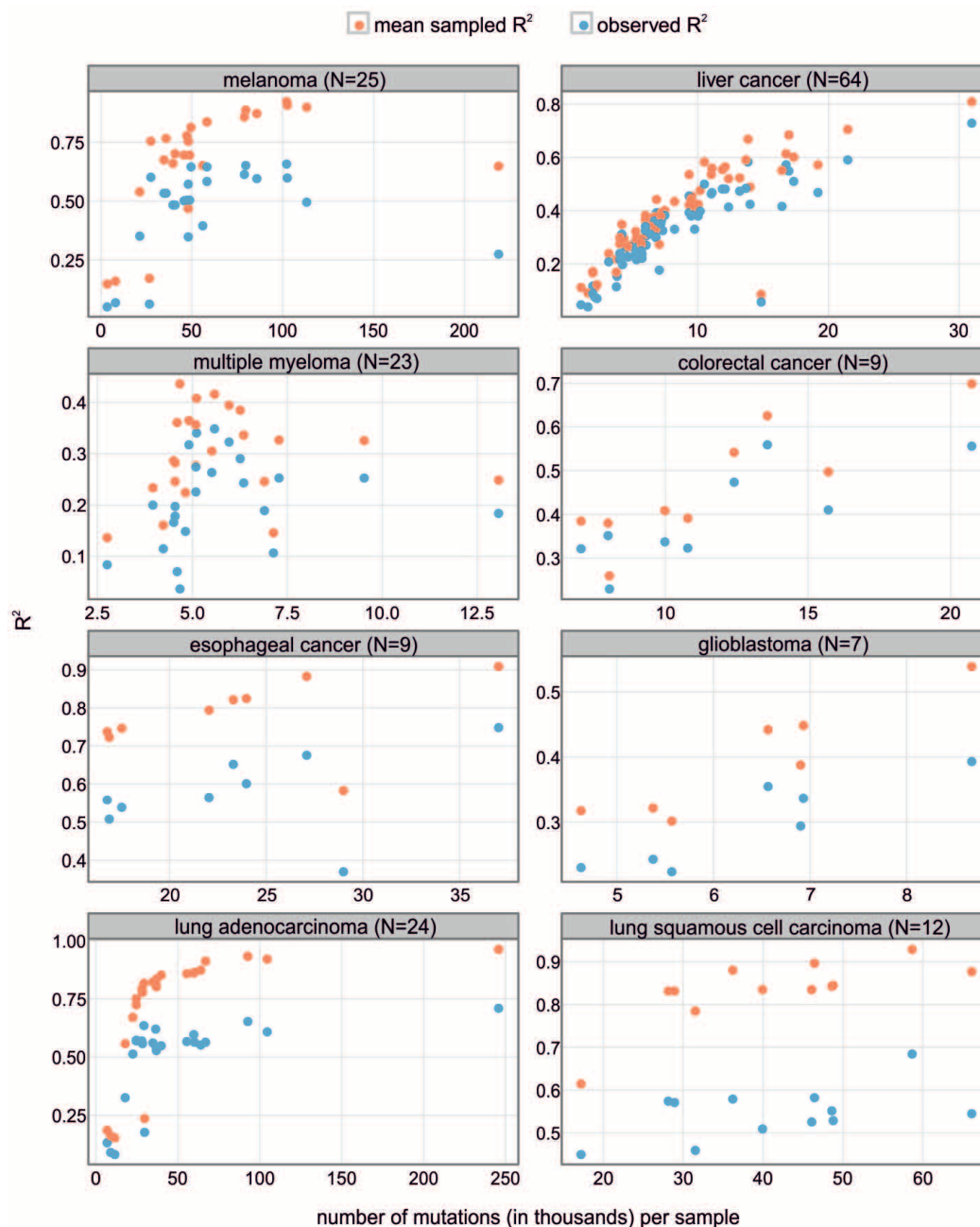


Extended Data Figure 3 | Scatter plots of the measured number of somatic mutations per Mb in different cancer genomes versus the number of mutations predicted by the Random Forest algorithm. The training set consisted of 10% of the data, the remaining 90% was used to test the predictions.



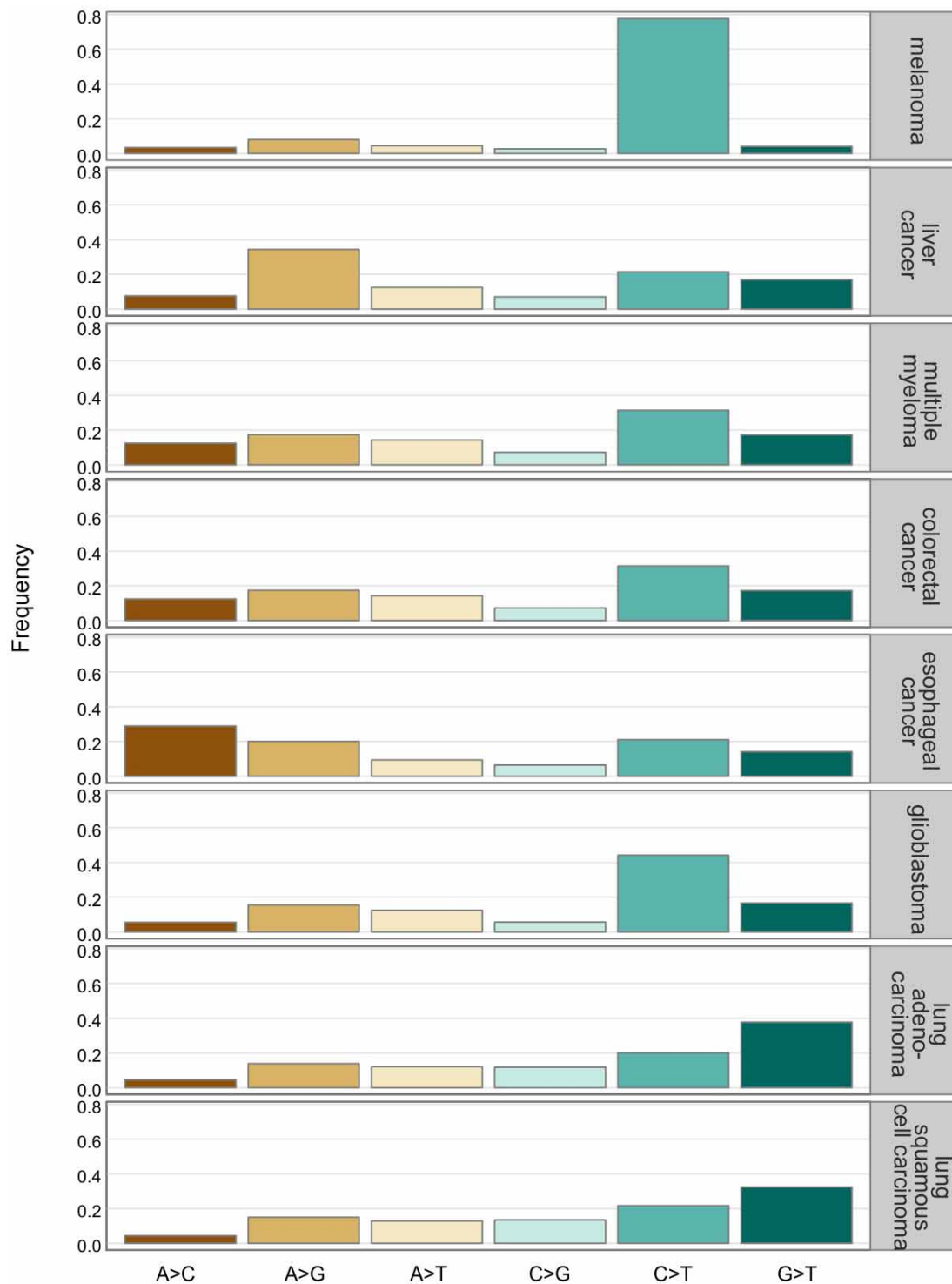
Extended Data Figure 4 | Prediction accuracy of the models trained on individual cancers as a function of the number of mutations. The red line represents the prediction accuracy of the model used to predict the mutation

density of samples pooled by cancer type (sum of all mutations in individual cancers of a certain cancer type). N – number of individual cancers per cancer type.

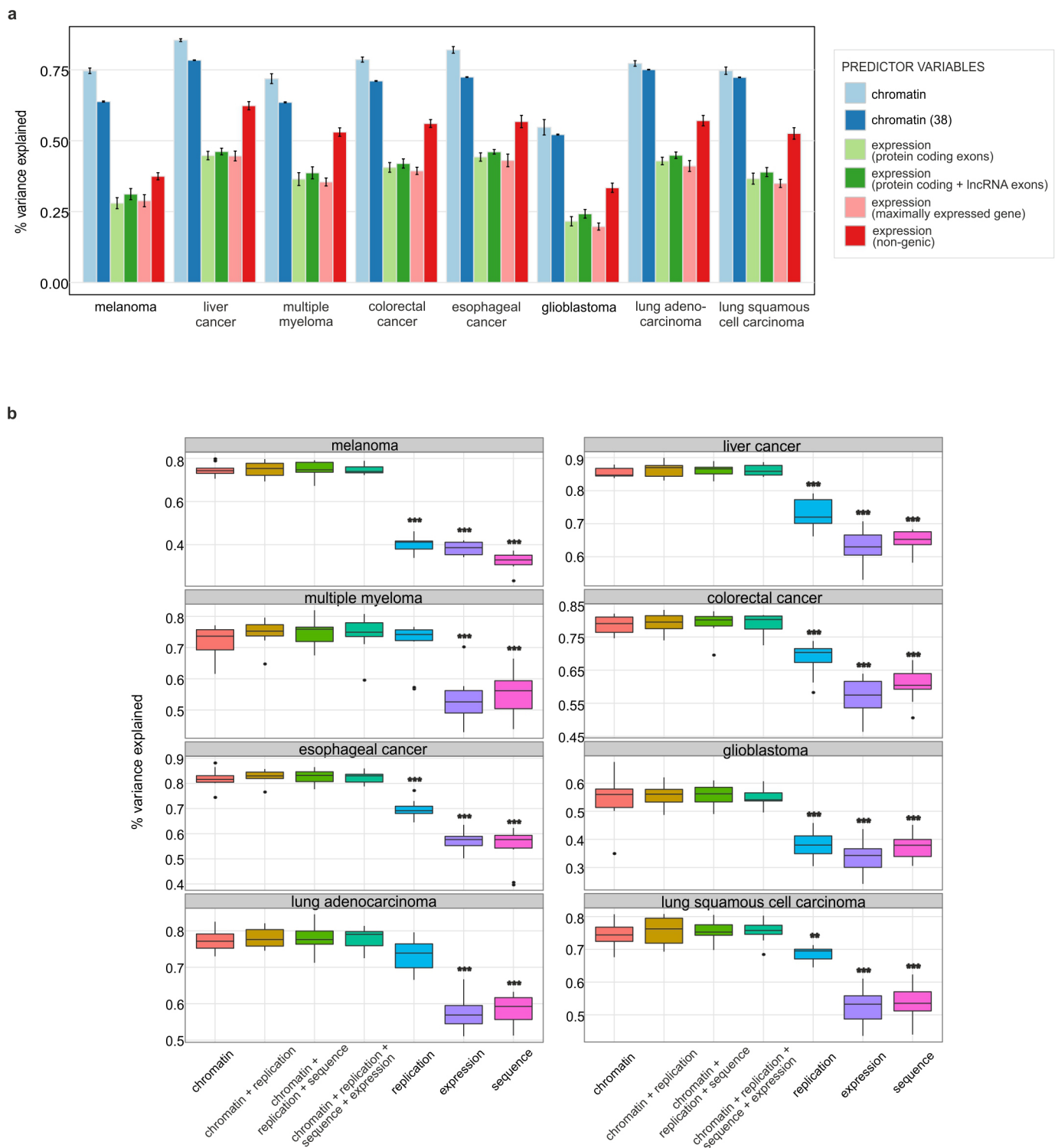


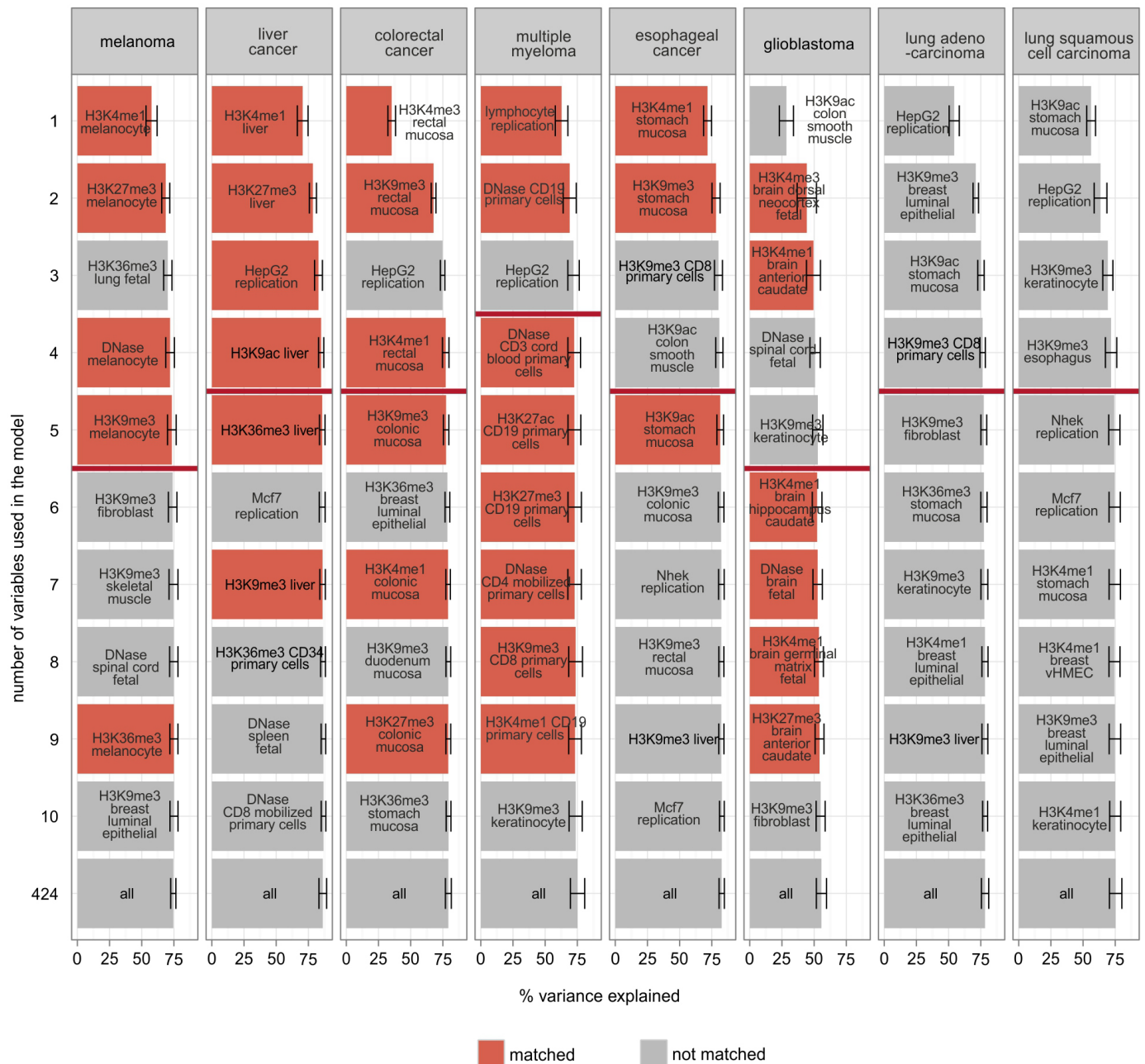
Extended Data Figure 5 | Sampling variance. Red, the squared correlation coefficient (R^2) between the observed mutational profile and the profiles predicted by Random Forest. Blue, the maximal attainable variance explained, calculated as the average correlation coefficient squared (R^2) between the

mutational profiles predicted by Random Forest and 100 simulated mutational profiles modelled as a Poisson distribution with the mean predicted by epigenomic features. N – number of individual cancers per cancer type.



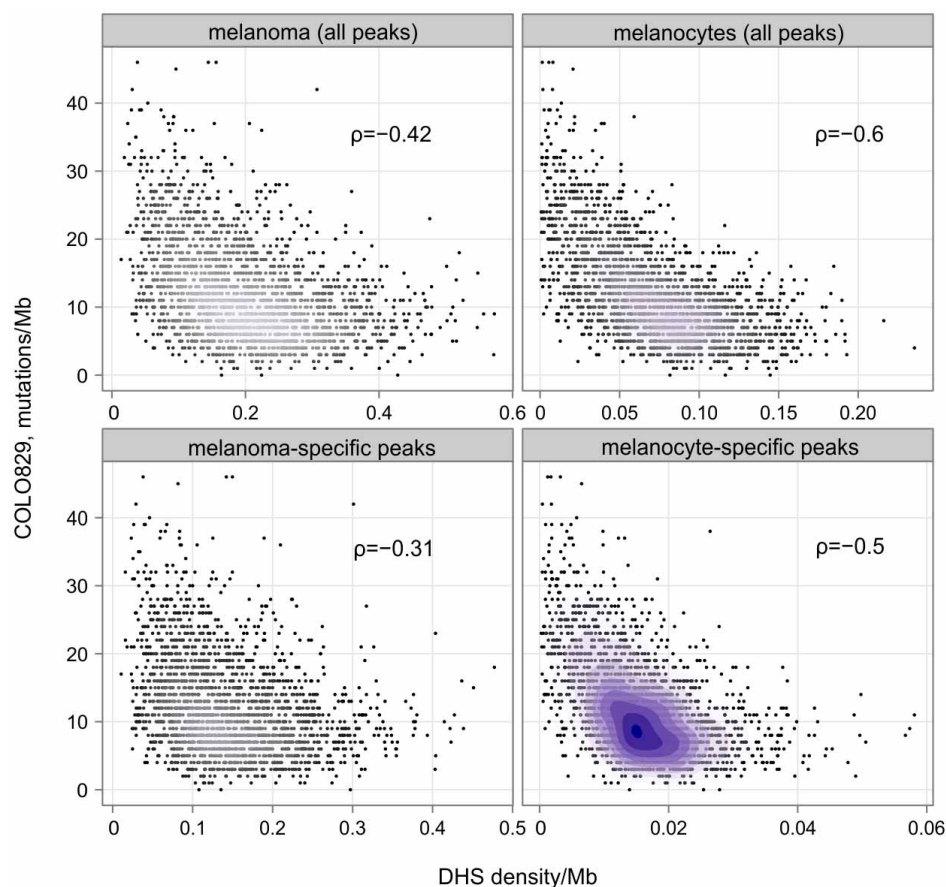
Extended Data Figure 6 | Frequency of different types of mutations in different cancer types.





Extended Data Figure 8 | Feature selection by using the backward elimination procedure. For each cancer type, variables are ordered from top to bottom by decreasing importance. Each bar represents the fraction of variance explained by the model using the corresponding bar and all bars above it. Error bars represent the standard error of the mean variance explained

by the model, estimated using tenfold cross-validation. The red line indicates the cutoff needed to achieve the prediction accuracy of the full model - 1 s.e.m. For each cancer type, features measured in related cell lines are shown in red.



Extended Data Figure 9 | The number of mutations per megabase in COLO829 cell line versus DHS density in melanoma cell lines, melanocytes, DHSs specific to melanomas and DHSs specific to melanocytes. Correlation is calculated using the Spearman's rank correlation coefficient. DHS density in melanoma cell lines corresponds to DHSs measured in 11

melanoma cell lines. DHSs specific to melanomas correspond to DHSs observed in melanomas but not observed in melanocytes. DHSs specific to melanocytes correspond to DHSs observed in melanocytes but not observed in melanomas.

Extended Data Table 1 | Cell types for which mRNA-seq data was downloaded from Epigenomics Roadmap

adrenal gland
bladder
brain fetal
brain germinal matrix fetal
brain hippocampus middle
breast luminal epithelial cells
breast myoepithelial cells
breast vHMEC
CD34 mobilized primary cells
esophagus
gastric
H1 BMP4 derived mesendoderm cultured cells
H1 BMP4 derived trophoblast cultured cells
H1 cell line
H1 derived mesenchymal stem cells
H1 derived neuronal progenitor cultured cells
IMR90 cell line
iPS DF 19.11 cell line
iPS DF 6.9 cell line
liver
lung
lung left fetal
lung right fetal
muscle arm fetal
muscle back fetal
muscle leg fetal
muscle trunk fetal
ovary
ovary fetal
pancreas
penis foreskin fibroblast primary cells
penis foreskin keratinocyte primary cells
penis foreskin melanocyte primary cells
psoas muscle
sigmoid colon
small intestine
spinal cord fetal
thymus

Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease

Elizabeta Gjoneska^{1,2*}, Andreas R. Pfenning^{2,3*}, Hansruedi Mathys¹, Gerald Quon^{2,3}, Anshul Kundaje^{2,3,4}, Li-Huei Tsai^{1,2§} & Manolis Kellis^{2,3§}

Alzheimer's disease (AD) is a severe¹ age-related neurodegenerative disorder characterized by accumulation of amyloid- β plaques and neurofibrillary tangles, synaptic and neuronal loss, and cognitive decline. Several genes have been implicated in AD, but chromatin state alterations during neurodegeneration remain uncharacterized. Here we profile transcriptional and chromatin state dynamics across early and late pathology in the hippocampus of an inducible mouse model of AD-like neurodegeneration. We find a coordinated down-regulation of synaptic plasticity genes and regulatory regions, and upregulation of immune response genes and regulatory regions, which are targeted by factors that belong to the ETS family of transcriptional regulators, including PU.1. Human regions orthologous to increasing-level enhancers show immune-cell-specific enhancer signatures as well as immune cell expression quantitative trait loci, while decreasing-level enhancer orthologues show fetal-brain-specific enhancer activity. Notably, AD-associated genetic variants are specifically enriched in increasing-level enhancer orthologues, implicating immune processes in AD predisposition. Indeed, increasing enhancers overlap known AD loci lacking protein-altering variants, and implicate additional loci that do not reach genome-wide significance. Our results reveal new insights into the mechanisms of neurodegeneration and establish the mouse as a useful model for functional studies of AD regulatory regions.

Gene expression^{2,3} and genetic variation⁴ studies suggest gene-regulatory changes may underlie AD, but regulatory epigenetic alterations during neurodegeneration remain uncharacterized, given the inaccessible nature of human brain samples. To address this need, we profiled transcriptional and epigenomic changes during neurodegeneration in the hippocampus of the CK-p25 mouse model of AD^{5–7} and CK littermate controls at both early and late stages of neurodegeneration (2 weeks and 6 weeks after p25 induction). CK-p25 mice, in which accumulation of the Cdk5 activator protein p25 is inducible, exhibit DNA damage, aberrant gene expression and increased amyloid- β levels at early stages⁷, followed by neuronal and synaptic loss and cognitive impairment at late stages^{5,6}.

For transcriptome analysis, we used RNA sequencing to quantify gene expression changes for 13,836 ENSEMBL genes (see Methods, Extended Data Fig. 1a and Supplementary Table 1). We found 2,815 upregulated genes and 2,310 downregulated genes in the CK-p25 AD mouse model as compared to CK littermate controls (at $q < 0.01$; Supplementary Table 1), which we classified into transient (2 weeks only), late-onset (6 weeks only) and consistent (both) expression classes (Fig. 1a, Extended Data Fig. 4a and Supplementary Table 1). These showed distinct functional enrichments (Fig. 1a and Supplementary Table 2), with transient-increase genes enriched in cell cycle functions ($P < 10^{-92}$), consistent-increase genes enriched in immune ($P < 10^{-10}$) and stimulus-response ($P < 10^{-4}$) functions, and consistent- and late-decrease genes enriched in synaptic and learning functions ($P < 10^{-12}$).



These coordinated neuronal and immune changes are consistent with the pathophysiology of AD² and probably

reflect both cell-type-specific expression changes and changes in cell composition. Indeed, comparison with expression in microglia⁸ (the resident immune cells of the brain) shows that both the cell type composition ($P = 2.7 \times 10^{-4}$) and microglia-specific activation ($P = 2.9 \times 10^{-6}$) significantly contribute to the gene expression changes (see Methods). Additionally, reverse transcription followed by quantitative PCR (RT-qPCR) of increased-level genes in purified CD11b⁺ CD45^{low} microglia populations confirms cell-type-specific activation for five of the seven microglia-specific genes tested (Extended Data Fig. 2).

Confirming the biological relevance of our mouse model for human AD, the observed changes in gene expression in mouse, especially for the consistent and late classes, agreed with gene expression differences between 22 patients with AD and 9 controls in human post-mortem laser capture microdissected hippocampal grey matter² (Fig. 1b). The enriched Gene Ontology classes also agreed between mouse and human, with higher immune gene expression and lower neuronal gene expression in patients with AD (Fig. 1c).

For epigenome analysis, we used chromatin immunoprecipitation sequencing (ChIP-seq) to profile seven chromatin marks⁹: histone 3 Lys 4 trimethylation (H3K4me3; associated primarily with active promoters); H3K4me1 (enhancers); H3K27 acetylation (H3K27ac; enhancer/promoter activation); H3K27me3 (Polycomb repression); H3K36me3 and H4K20me1 (transcription); and H3K9me3 (heterochromatin) (Extended Data Fig. 1a). We used ChromHMM (<http://compbio.mit.edu/ChromHMM/>) to learn a chromatin state model (Methods and Extended Data Fig. 3a) defined by recurrent combinations of histone modifications, consisting of promoters, enhancers, transcribed, bivalent, repressed, heterochromatin and low-signal states (Extended Data Fig. 3a). We defined 57,840 active promoters using H3K4me3 peaks within promoter chromatin states, and 151,447 active enhancer regions using H3K27ac peaks within enhancer chromatin states (Extended Data Fig. 1a, Supplementary Table 3 and Methods).

We mapped orthologous genes between mouse and human using ENSEMBL one-to-one orthologues (see Methods). We also mapped orthologous noncoding regions using multiple mammalian sequence alignments, mapping each mouse peak to its best human match (see Methods). We found matches for 90% of promoter regions, 84% of enhancers, 74% of Polycomb-repressed regions and 33% of heterochromatin regions (Supplementary Table 3). Comparing our mouse chromatin states to human hippocampus chromatin states¹⁰, we found significant epigenomic conservation at orthologous noncoding regions (Extended Data Fig. 3b), consistent with recent results¹¹.

¹The Picower Institute for Learning and Memory, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. ²Broad Institute of Harvard University and Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, USA. ³Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. ⁴Department of Genetics, Department of Computer Science, Stanford University, Stanford, California 94305, USA.

*These authors contributed equally to this work.

§These authors jointly supervised this work.

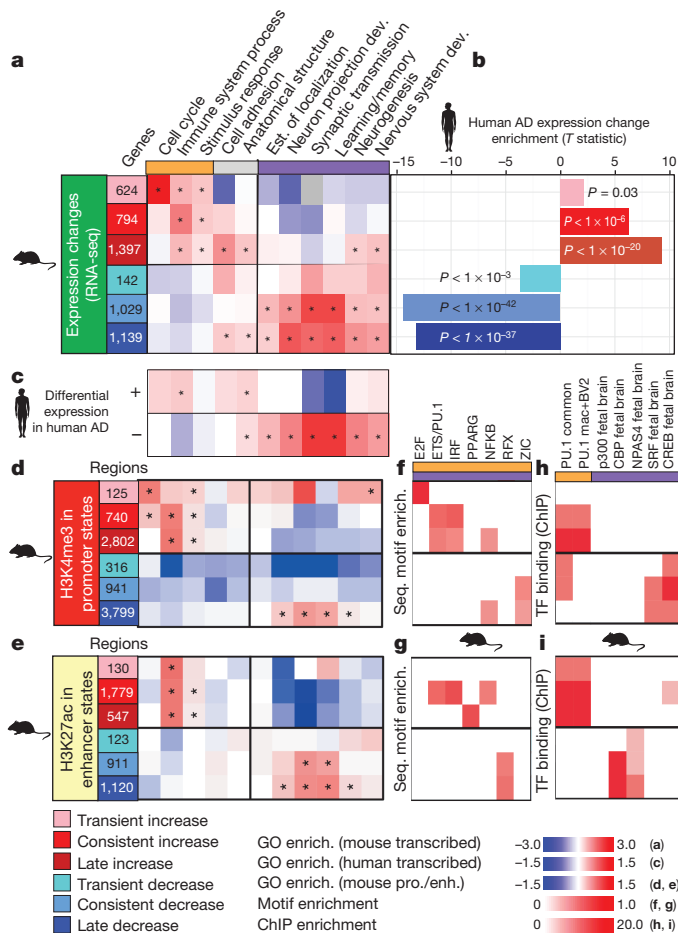


Figure 1 | Conserved gene expression changes between mouse and human AD are associated with immune and neuronal functions. **a**, Six distinct temporal classes of differentially expressed genes are denoted; transient (early) increase (pink) or decrease (light blue), consistent increase (red) or decrease (blue), and late (6 week) increase (dark red) or decrease (navy blue). Expression is shown relative to the mean of three replicates at 2-week control (CK) mice. Shown are the most significant distinct biological process Gene Ontology (GO) categories in each class of differentially regulated genes (asterisk denotes enrichment of hypergeometric $P < 0.01$). Grey boxes indicate no overlapping genes. **b**, T -statistic identifying the bias of each differentially regulated class of genes in AD cases relative to controls; negative t denotes lower expression in AD, positive t denotes higher expression in AD. **c**, Enrichment of Gene Ontology categories for differentially expressed genes between AD cases and controls in human. **d**, **e**, Enrichment of each Gene Ontology category examined in the gene expression analysis was calculated for H3K4me3 promoters (pro.; red) (d) and H3K27ac enhancers (enh.; yellow) (e). Asterisk denotes categories with a binomial $P < 0.01$. **f**, **g**, Enrichment of regulatory motifs within changing promoters (top) (f) and enhancers (bottom) (g) in the mouse AD model. **h**, **i**, Overlap of changing promoters (top) (h) and enhancers (bottom) (i) with regions shown to be bound by immune (orange) and neuronal (purple) transcriptional factors (TF) and co-factors profiled using ChIP-seq in mouse immune and neuronal tissues^{15–19}.

We quantified epigenomic changes in promoter regions using relative differences in H3K4me3 levels resulting in 3,667 increased-level and 5,056 decreased-level peaks ($q < 0.01$; Extended Data Fig. 4b and Supplementary Table 3), which we classified into transient, consistent and late-stage, as for gene expression changes. For enhancer regions, we used relative levels of H3K27ac, resulting in 2,456 increased-level and 2,154 decreased-level peaks (Extended Data Fig. 4c and Supplementary Table 3). Only a very small number of peaks showed differences in Polycomb-repressed and heterochromatin regions, leading us to focus on enhancer and promoter changes for the remaining analyses (Extended Data Fig. 4d, e and Supplementary Table 3).

Genes flanking increased- and decreased-level regulatory regions (see Methods) showed consistent gene expression changes for both promoter and enhancers regions (Extended Data Fig. 5), and were consistently enriched in immune and stimulus-response functions for increased-level enhancers and promoters, and in synapse and learning-associated functions for decreased-level enhancers and promoters (Fig. 1d, e), consistent with our Gene Ontology results of changing gene expression levels.

Increased- and decreased-level regulatory regions showed distinct regulatory motif enrichments (Fig. 1f, g). Increased-level peaks were enriched in NFkB, E2F, PPARG, IRF and PU.1 (ref. 12) transcription factor motifs for both enhancers and promoters, consistent with immune regulator targeting. Decreased-level peaks in enhancers were enriched for DNA-binding RFX motifs, and peaks in promoters were enriched for zinc-finger ZIC motifs, two known neurodevelopmental regulators^{13,14}.

Consistent with the observed motif enrichments, increased-level enhancers and promoters showed *in vivo* binding of PU.1 in mouse embryos^{15,16} (Fig. 1h, i). Only increased-level promoters were bound in macrophages and BV-2 microglial-like cells^{17–19} that are both implicated in AD²⁰, while both increased- and decreased-level promoters were bound in several immune cell lineages (Fig. 1h). The PU.1 regulator itself (encoded by the *SPI1* gene) showed increased expression and enhancer levels (Extended Data Fig. 1b), possibly contributing to immune enhancer and promoter upregulation, consistent with roles for PU.1, ETS-1 and other ETS family members in microglia activation and proliferation during neurodegeneration^{21,22}. By contrast, neuronal function regulators were not enriched in increased-level enhancers (except for a weak enrichment of fetal brain CREB; Fig. 1i), consistent with primarily immune and inflammatory function of these regions.

Decreased-level enhancers and promoters were targeted by different regulators, suggesting distinct regulatory programs. Decreased-level promoters were preferentially bound by CREB and SRF ($P < 10^{-21}$ and $P < 10^{-16}$), two known regulators of neuronal activity in cortical neurons²³, and decreased-level enhancers were preferentially bound by CBP ($P_{\text{hypergeometric}} = 5.4 \times 10^{-20}$), a known co-activator for neuronal activity¹⁶ (Fig. 1h, i). Surprisingly, p300-bound regions¹⁵ did not show any enrichment, suggesting distinct roles for CBP and p300, despite a general association with enhancers for both. The distinct neuronal and immune targeting of decreased-level and increased-level regulatory regions provides a mechanistic basis for the expression differences observed for neuronal and immune genes, and suggests potential therapeutic targets for reversing observed alterations during neurodegeneration.

On the basis of chromatin state annotations in 127 human cell types and tissues¹⁰ (Fig. 3a and Supplementary Table 4), regions orthologous to increased-level enhancers in mouse showed immune cell enhancer activity in human ($P < 10^{-100}$), while orthologues of decreased-level enhancers in mouse showed fetal brain tissue enhancer activity in human ($P < 10^{-8}$ consistent; $P < 10^{-17}$ late-stage; Fig. 2a and Supplementary Table 4). Adult brain tissues (including hippocampus) were not as strongly enriched, suggesting changes are biased towards neuronal plasticity. These results are consistent with decreased neuronal plasticity, and increased microglial activation and proliferation during AD progression²⁴.

To verify whether the increased-level putative enhancer regions were indeed functional, we used a luciferase reporter assay to evaluate their ability to drive *in vitro* gene expression in immortalized murine microglial (BV-2) and neuroblastoma (N2a) cell lines. Eight of the nine increased-level human orthologues tested were indeed able to drive *in vitro* reporter expression. Two of these, BIN1 and ZNF710, were active in both cell types, while the remaining six showed a BV-2-cell-specific increase in luciferase expression (Fig. 2b and Supplementary Table 5), confirming both functional conservation and tissue specificity of increased-level enhancer regions implicated by our mouse model of AD.

Human orthologues of increased-level enhancers were also enriched for expression quantitative trait loci (eQTLs) in CD4⁺ T cells and CD14⁺ monocytes^{25,26} (Extended Data Fig. 6 and Supplementary Table 6),

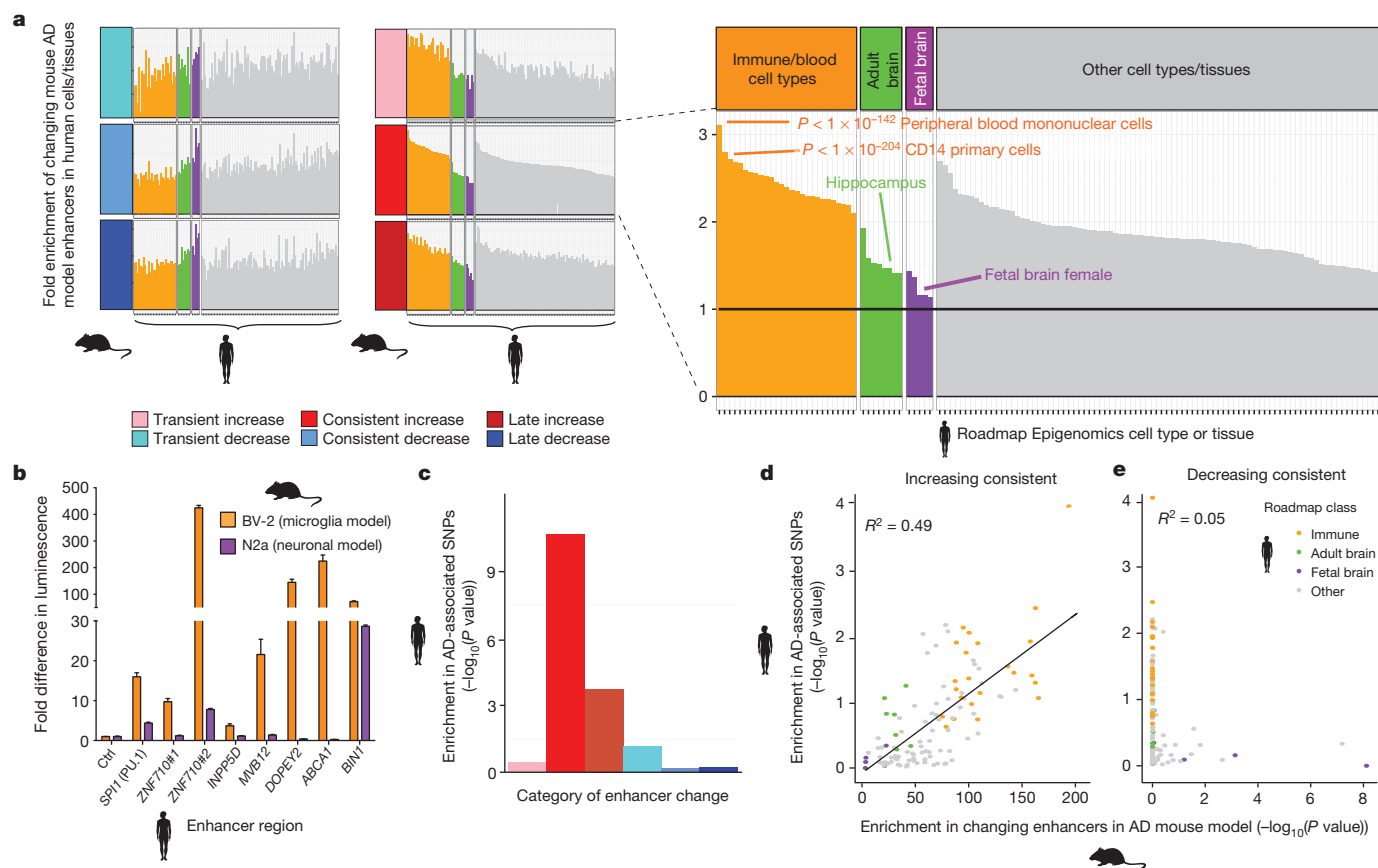


Figure 2 | AD GWAS loci are preferentially enriched in increasing enhancer orthologues with immune function. **a**, Enrichment (y axis) of changing mouse AD enhancer orthologues, with a focus on consistently increasing (red) category of enhancers, in 127 cell and tissue types profiled by the Roadmap Epigenomics Consortium¹⁰ (columns). Roadmap samples are grouped into fetal brain (purple), adult brain (green), immune/blood cell types (orange) and all other (grey). **b**, Cell-type-specific fold luciferase reporter expression change relative to control (ctrl) for selected increasing enhancer regions in BV-2 microglia (orange) versus N2A neurons (purple) ($n = 3$, $*P < 0.05$, two-tailed

t -test). **c**, Enrichment of AD-associated SNPs (y axis, binomial P value) in human regions orthologous to the mouse enhancers. **d**, **e**, Enrichment of AD-associated SNPs (y axis, permutation P value) in tissue-specific enhancer annotations from the Roadmap Epigenomics Consortium (points), relative to their enrichment for consistently increasing (**d**) and consistently decreasing (**e**) orthologous enhancer regions in the mouse AD model (x axis, hypergeometric P value). Linear regression trend line and R^2 , based on Pearson correlation, is shown.

indicating that they contain driver mutations controlling immune cell regulatory programs. The enrichment was strongest for CD14⁺ monocytes (Extended Data Fig. 6), which also showed the highest enhancer enrichment and is consistent with the observed inflammatory response Gene Ontology category.

To test whether the implicated regulatory regions are causal, we examined their enrichment for AD-associated variants from genome-wide association studies (GWAS). Genetic variants associated with AD in a meta-analysis of ~74,000 individuals⁴ were enriched in increased-level enhancer orthologues (Fig. 2c) (4.4-fold enrichment, binomial $P = 1.2 \times 10^{-10}$ at GWAS cutoff $P < 0.001$; 9.7-fold enrichment, binomial $P < 3.7 \times 10^{-6}$ at GWAS cutoff $P < 10^{-5}$). By contrast, decreased-level enhancer orthologues were surprisingly not enriched (0.61-fold), suggesting a causal role specifically for immune-related processes. Promoter regions were only weakly enriched, strongly implicating distal enhancers in mediating AD predisposition (Extended Data Fig. 7).

Across diverse cell types and tissues, we found concordance between the enrichment for AD GWAS single nucleotide polymorphisms (SNPs) and the enrichment for increased-level enhancer orthologues ($R^2 = 0.49$; Fig. 2d, Extended Data Fig. 8a, left and Supplementary Table 4), with CD14⁺ immune cells being the most enriched in both, followed by other immune cell types, and with fetal brain enhancers showing the smallest enrichment in both. By contrast, decreasing enhancers orthologues showed a very weak correlation ($R^2 < 0.08$) (Fig. 2e, Extended Data Fig. 8b, right and Supplementary Table 4). The increased-level enhancer orthologue

enrichment for AD GWAS SNPs persisted both within CD14⁺ enhancers (3.0-fold enrichment, binomial $P = 1.3 \times 10^{-5}$) and outside CD14⁺ enhancers (3.4-fold, $P = 0.005$), suggesting it is not solely a feature of CD14⁺ cell type enrichment (see Methods).

These results are consistent with enhanced microglial expression of CD14 in brains of animal models of AD, and a regulatory role of the CD14 receptor in microglial inflammatory response, which modulates amyloid- β deposition²⁴. Thus, the enrichment of AD-associated variants in CD14⁺ primary immune cells, but not neuronal cells, indicates that AD genetic predisposition is primarily associated with immune function, while decrease in neuronal plasticity may be affected primarily by non-genetic effects, such as diet, education, physical activity and age, which are thought to lead to epigenetic changes related to cognitive reserve²⁷.

We next used the epigenomic annotations of increased-level enhancer orthologues to gain insights into AD-associated loci (Supplementary Table 7). Among the 20 genome-wide significant AD-associated loci⁴, 11 contain no protein-altering SNPs in linkage disequilibrium (LD), indicating they may have noncoding roles. Of these, five localize within increased-level enhancer orthologues, including two well-established GWAS loci (*PICALM* and *BIN1*), and three loci (*INPP5D*, *CELF1* (also containing the *SPI1* gene) and *PTK2B*) only recently recognized as significant by combining all AD cohorts.

For *INPP5D* (Fig. 3a), a known regulator of inflammation²⁸, the most significant variants localize within an increased-level enhancer orthologue, which also shows CD14⁺ enhancer activity. In the *CELF1* locus

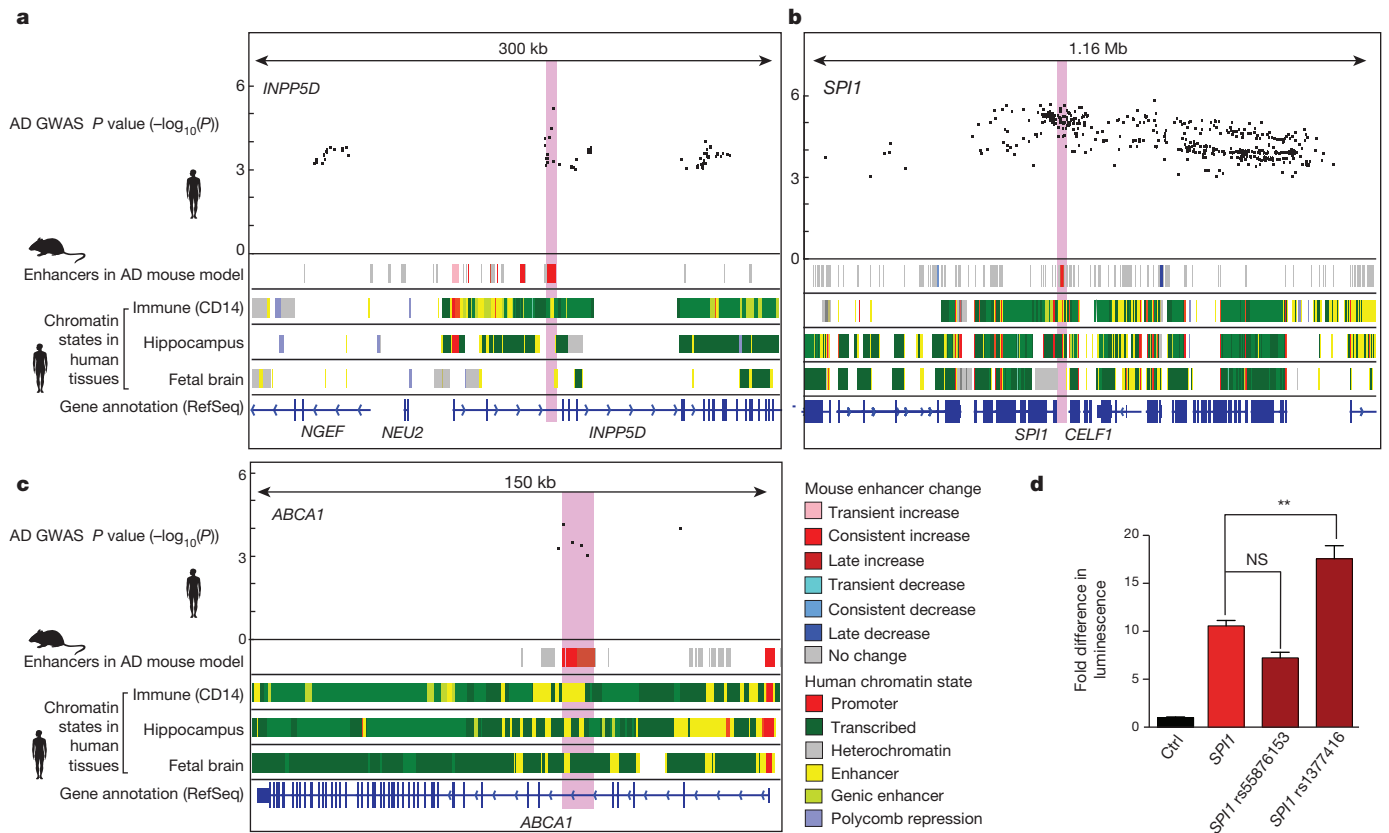


Figure 3 | Increasing enhancer orthologues help interpret AD-associated non-coding loci. **a–c**, Overlap of disease-associated SNPs (top) with increasing enhancers (second row, red) and immune enhancers in human (CD14⁺ primary cells) is shown for genome-wide significant (*INPP5D* and *CELF1* (containing the *SPI1* gene); **a** and **b**) and below-significance (*ABCA1*; **c**) AD GWAS loci. Roadmap chromatin state annotations for immune cells (CD14⁺ primary; E029), hippocampus (E071) and fetal brain (E81), with colours as

shown in the key. Light red highlight denotes increasing enhancer regions tested in luciferase assay. kb, kilobases; Mb, megabases. **d**, AD-associated SNP rs1377416 amplifies *in vitro* luciferase activity of putative enhancer region 38,313–37,359 base pairs (bp) upstream of *SPI1* (PU.1) gene in BV-2 cells. $n = 3$, $P < 0.0001$, one-way analysis of variance (ANOVA); ** $P < 0.01$, Tukey's multiple comparison post-hoc test. NS, not significant.

(Fig. 3b) a large region of association spans several genes, but the strongest genetic signal ($P = 2 \times 10^{-6}$) localizes upstream of *SPI1* (PU.1), and specifically within an increased-level enhancer orthologue that is also active in immune cells. We confirmed that the AD-associated C–T substitution, rs1377416, in the *SPI1* enhancer leads to increased *in vitro* enhancer activity in murine BV-2 microglia cells using a luciferase reporter assay (Fig. 3d). In addition, the AD-associated SNP rs55876153 near *SPI1*, which overlaps an increased-level mouse enhancer orthologue, is in strong linkage disequilibrium ($LD = 0.89$, see Methods) with a known *SPI1* eQTL, rs10838698 (ref. 25), even though it did not significantly alter enhancer activity in the luciferase assay.

Outside known GWAS loci, an additional 22 weakly associated regions (3.9-fold, $P < 4.9 \times 10^{-7}$) contain variants within increased-level enhancer orthologues (Supplementary Table 7), of which 17 lack protein-altering variants in linkage disequilibrium ($R^2 < 0.4$), providing strong candidates for directed experiments. One such example includes *ABCA1* ($P = 6.9 \times 10^{-5}$; Fig. 3c), a paralogue of AD-associated *ABCA7* and encoding a glial-expressed transporter that influences APOE metabolism in the central nervous system²⁹. The region lacks protein-altering variants and all five SNPs in the cluster of association lie specifically within an increased-enhancer orthologue, which is also active in CD14⁺ immune cells and, to a lesser extent, in human hippocampus and fetal brain.

Overall, our study revealed contrasting changes in immune and neuronal genes and regulatory regions during AD-like neurodegeneration in mouse, strong human–mouse conservation of gene expression and epigenomic signatures, and enrichment of AD-associated loci in increased-level enhancer orthologues in human. While immune genes are known

to be among the most significant genetic loci associated with AD, the depletion of neuronal promoters and enhancers is particularly notable for a cognitive disorder with well-established environmental and experiential factors that include diet, exercise, education and age. These results are consistent with a model in which increased immune susceptibility to environmental factors during ageing and cognitive decline is mediated by interactions between genetically driven immune cell dysregulation and environmentally driven epigenomic alteration in neuronal cells.

Our study also illustrates the power of model organisms for the study of human disease progression, especially for disorders affecting inaccessible tissues for which only post-mortem samples are available in human. We find that molecular changes in both genes and regulatory regions are highly conserved between human AD and CK-p25 neurodegeneration, enabling detailed studies of the molecular signatures associated with disease progression across diverse environmental conditions, in a variety of brain regions and cell types, and in response to therapeutic agents before or after disease onset.

Lastly, our results indicate specific therapeutic targets for AD, including putative causal nucleotides lying in increased-level enhancer orthologues that may be targeted by CRISPR/Cas9 genome editing³⁰, and *trans*-acting regulators. In particular, the transcription factor PU.1 is implicated as a therapeutic target by its genetic association with AD, as well as the enrichment of the PU.1 motif and the PU.1 *in vivo* binding sites at increased-level regulatory regions during mouse neurodegeneration. The conservation of neuronal and immune regulatory circuitry between mouse and human suggests that CK-p25 mice may offer a powerful model for studying the gene-regulatory and cognitive effects of such interventions.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 7 January 2014; accepted 22 January 2015.

- Alzheimer's Association. 2014 Alzheimer's disease facts and figures. *Alzheimers Dement* **10**, e47–e92 (2014).
- Blalock, E. M., Buechel, H. M., Popovic, J., Geddes, J. W. & Landfield, P. W. Microarray analyses of laser-captured hippocampus reveal distinct gray and white matter signatures associated with incipient Alzheimer's disease. *J. Chem. Neuroanat.* **42**, 118–126 (2011).
- Zhang, B. *et al.* Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. *Cell* **153**, 707–720 (2013).
- Lambert, J. C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature Genet.* **45**, 1452–1458 (2013).
- Cruz, J. C., Tseng, H.-C., Goldman, J. A., Shih, H. & Tsai, L.-H. Aberrant Cdk5 activation by p25 triggers pathological events leading to neurodegeneration and neurofibrillary tangles. *Neuron* **40**, 471–483 (2003).
- Fischer, A., Sananbenesi, F., Pang, P. T., Lu, B. & Tsai, L.-H. Opposing roles of transient and prolonged expression of p25 in synaptic plasticity and hippocampus-dependent memory. *Neuron* **48**, 825–838 (2005).
- Cruz, J. C. *et al.* p25/cyclin-dependent kinase 5 induces production and intraneuronal accumulation of amyloid beta *in vivo*. *J. Neurosci.* **26**, 10536–10541 (2006).
- Orre, M. *et al.* Isolation of glia from Alzheimer's mice reveals inflammation and dysfunction. *Neurobiol. Aging* **35**, 2746–2760 (2014).
- ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
- Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* <http://dx.doi.org/nature14248> (this issue).
- Cheng, Y. *et al.* Principles of regulatory information conservation between mouse and human. *Nature* **515**, 371–375 (2014).
- Gallant, S. & Gilkeson, G. ETS transcription factors and regulation of immunity. *Arch. Immunol. Ther. Exp. (Warsz.)* **54**, 149–163 (2006).
- Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl Acad. Sci. USA* **107**, 21931–21936 (2010).
- Aruga, J. The role of *Zic* genes in neural development. *Mol. Cell. Neurosci.* **26**, 205–221 (2004).
- Visel, A. *et al.* A high-resolution enhancer atlas of the developing telencephalon. *Cell* **152**, 895–908 (2013).
- Kim, T. K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**, 182–187 (2010).
- May, G. *et al.* Dynamic analysis of gene expression and genome-wide transcription factor binding during lineage specification of multipotent progenitors. *Cell Stem Cell* **13**, 754–768 (2013).
- Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
- Crotti, A. *et al.* Mutant Huntingtin promotes autonomous microglia activation via myeloid lineage-determining factors. *Nature Neurosci.* **17**, 513–521 (2014).
- Prinz, M. & Priller, J. Microglia and brain macrophages in the molecular age: from origin to neuropsychiatric disease. *Nature Rev. Neurosci.* **15**, 300–312 (2014).
- Gómez-Nicola, D., Fransen, N. L., Suzzi, S. & Perry, V. H. Regulation of microglial proliferation during chronic neurodegeneration. *J. Neurosci.* **33**, 2481–2493 (2013).
- Jantarantotai, N. *et al.* Upregulation and expression patterns of the angiogenic transcription factor Ets-1 in Alzheimer's disease brain. *J. Alzheimers Dis.* **37**, 367–377 (2013).
- Lyons, M. R. & West, A. E. Mechanisms of specificity in neuronal activity-regulated gene transcription. *Prog. Neurobiol.* **94**, 259–295 (2011).
- Reed-Geaghan, E. G., Reed, Q. W., Cramer, P. E. & Landreth, G. E. Deletion of CD14 attenuates Alzheimer's disease pathology by influencing the brain's inflammatory milieu. *J. Neurosci.* **30**, 15369–15373 (2010).
- Fairfax, B. P. *et al.* Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science* **343**, 1246949 (2014).
- Raj, T. *et al.* Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. *Science* **344**, 519–523 (2014).
- Stern, Y. Cognitive reserve in ageing and Alzheimer's disease. *Lancet Neurol.* **11**, 1006–1012 (2012).
- Lam, P. Y., Yoo, S. K., Green, J. M. & Huttenlocher, A. The SH2-domain-containing inositol 5-phosphatase (SHIP) limits the motility of neutrophils and their recruitment to wounds in zebrafish. *J. Cell Sci.* **125**, 4973–4978 (2012).
- Krimbou, L. *et al.* Molecular interactions between apoE and ABCA1: impact on apoE lipidation. *J. Lipid Res.* **45**, 839–848 (2004).
- Ran, F. A. *et al.* Genome engineering using the CRISPR-Cas9 system. *Nature Protocols* **8**, 2281–2308 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank A. Mungenast for critical reading and editing of the manuscript and discussions about the project, M. Taylor for mouse colony maintenance, and X. Zhang, R. Issner, H. Whitton and C. Epstein for technical assistance with ChIP-seq library preparation. We thank P. Kheradpour for the transcription factor binding site motif scan of the mouse genome. This work was partially supported by the Belfer Neurodegeneration Consortium funding and NIH/NINDS/NIA (R01NS078839) to L.-H.T., Early Postdoc Mobility fellowship from the Swiss National Science Foundation (P2BSP3_151885) to H.M., and NIH/NHGRI (R01HG004037-07 and RC1HG005334) to M.K.

Author Contributions This study was designed by E.G., A.R.P., A.K., M.K. and L.-H.T., and directed and coordinated by M.K. and L.-H.T. E.G. initiated, planned and performed the experimental work. A.R.P. performed computational analysis to characterize differential gene expression and histone mark levels, identify orthologous human regions and enriched transcription factor binding sites, and compare regulatory regions to human AD meta-analysis data. A.K. contributed to the computational analysis by generating mouse chromatin states and the quantification and control of ChIP datasets. H.M. helped with isolation and gene expression analysis of specific cell type populations. G.Q. performed permutation test comparing human Roadmap enhancers to AD GWAS SNPs. The manuscript was written by E.G., A.R.P., L.-H.T. and M.K., and commented on by all authors.

Author Information All data are available from the NCBI Gene Expression Omnibus (GEO) database under accession number GSE65159, the NIH Roadmap (<http://www.roadmapepigenomics.org/data>) and NCBI Epigenomics portal (<http://www.ncbi.nlm.nih.gov/epigenomics>). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to L.-H.T. (lhstai@mit.edu) or M.K. (manoli@mit.edu).



This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported licence. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons licence, users will need to obtain permission from the licence holder to reproduce the material. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-sa/3.0>

METHODS

Animals. All mouse work was approved by the Committee on Animal Care of the Division of Comparative Medicine at MIT. Adult (3-month-old) female double-transgenic CK-p25 (ref. 5) mice and their respective control littermates were used for the experiments. Brain tissue was collected at either 2 or 6 weeks after p25 induction. Upon dissection tissue was flash-frozen in liquid nitrogen. No animals were excluded from the study and no randomization or blinding was required.

Chromatin immunoprecipitation. Mouse hippocampus was collected immediately after euthanasia. Chromatin immunoprecipitation was then performed as described in Broad ChIP protocol (<http://www.roadmapepigenomics.org/protocols/type/experimental/>). In brief, tissues were minced and crosslinked in 1% formaldehyde (Thermo Scientific) for 15 min at room temperature and quenched with glycine for 5 min (Sigma). The samples were homogenized in cell lysis buffer containing protease inhibitors (complete, Roche) and chromatin was then fragmented to a size range of ~200–500 bp using a Branson 250 digital sonifier. Solubilized chromatin was then diluted and incubated with ~1 µg antibody at 4 °C overnight. Immune complexes were captured with Protein-A-sepharose beads, washed and eluted. Enriched chromatin was then subjected to crosslink reversal and proteinase K digestion at 65 °C, phenol–chloroform extraction and ethanol precipitation. Isolated ChIP DNA was resuspended and quantified using the Qubit assay (Invitrogen). H3K4me1 (Abcam, ab8895), H3K4me3 (Millipore, 07-473), H3K9me3 (Abcam, ab8898), H3K27me3 (Millipore, 07-449), H3K27ac (Abcam, ab4729), H3K36me3 (Abcam, ab9050) and H4K20me1 (Abcam, ab9051) were used to immunoprecipitate endogenous proteins. **ChIP-seq high-throughput sequencing, read mapping and quality control.** Sequencing libraries were prepared from ~1–5 ng ChIP (or input) DNA as described previously³¹. Gel electrophoresis was used to retain library fragments between 300 and 600 bp. Before sequencing, libraries were quantified using Qubit (Invitrogen) and quality-controlled using Agilent's Bioanalyzer. The 36-bp single-end sequencing was performed using the Illumina HiSeq 2000 platform according to standard operating procedures. For each histone modification, five biological replicate data sets were produced with corresponding whole-cell extract controls, except for H3K4me3, H4K20me1 and H3K27me3 in the 2-week control (CK) sample, where optimal amount of reads for sufficient coverage was obtained from four biological replicates. Reads were mapped to the mm9 reference mouse genome using MAQ v0.7.1-9 using default parameters³². Reads mapping to multiple locations were discarded. Duplicates were marked and filtered using PICARD (<http://picard.sourceforge.net/>). After filtering, roughly 55–60 million unique reads were obtained for each histone modification in each condition (~9–12 million reads per replicate) and ~110–145 million reads in total for the whole-cell extract controls in each condition. All replicate data sets passed quality control based on ENCODE ChIP-seq data standards based on read quality, read mapping statistics, library complexity and strand cross-correlation analysis (to measure signal-to-noise ratios)³³.

RNA sequencing. Mouse brains were homogenized and total RNA was extracted using Trizol reagent (Ambion). Total RNA was quality-controlled using Agilent's Bioanalyzer and prepared for sequencing using Illumina's TruSeq Stranded Total RNA Sample Preparation Kit with Ribo-Zero. High-throughput sequencing was performed on an Illumina HiSeq 2000 platform. Roughly 15 million 76-pair-end reads were generated for each data set. Sequence reads were aligned to mouse mm9 genome with Bowtie. On the basis of the reproducibility of the results (Fig. 2a), three replicate biological data sets were produced for each condition. A small number of replicates suffice for RNA sequencing (RNA-seq) studies³⁴ and we were able to detect large-scale changes in read counts in coherent gene ontology categories, with similarities to human AD (Fig. 2c, d). Therefore, we decided that additional replicates were not necessary.

Peak calling and signal coverage tracks for ChIP-seq data. For each histone modification in each condition, mapped reads were pooled across ChIP-seq replicates and regions of enrichment (peaks) were identified for the pooled ChIP-seq data set relative to the pooled control using the MACS2 peak caller (version 2.0.10.20130712)³⁵ (<https://github.com/taoliu/MACS/>) using a relaxed *p*-value of 0.01. For each histone modification, overlapping peaks (at least 1 bp overlap) were merged across all conditions to obtain a non-redundant master list of regions of enrichment. Master lists of broad domains of enrichment for the diffused marks H3K27me3, H3K9me3, H3K36me3 and H4K20me1 were obtained by allowing merging peaks across conditions that were within 1 kb of each other. Genome-wide signal coverage tracks representing per-base fold enrichment and the likelihood ratio of ChIP relative to control were also computed using MACS2.

Learning combinatorial chromatin states. We used ChromHMM to learn combinatorial chromatin states jointly across all four conditions³⁶. ChromHMM was trained using all seven chromatin marks in virtual concatenation mode across all conditions. Reads from replicate data sets were pooled before learning states. The ChromHMM parameters used are as follows: reads were shifted in the 5' to 3' direction by 100 bp; for each ChIP-seq data set, read counts were computed in non-overlapping 200-bp bins across the entire genome; each bin was discretized into

two levels, 1 indicating enrichment, and 0 indicating no enrichment. The binarization was performed by comparing ChIP-seq read counts to corresponding whole-cell extract control read counts within each bin and using a Poisson *P* value threshold of 1×10^{-4} (the default discretization threshold in ChromHMM). We trained several models with the number of states ranging from 12 to 23 states. We decided to use a 14-state model for all further analyses as it captured all the key interactions between the chromatin marks and larger number of states did not capture significantly new interactions. To assign biologically meaningful mnemonics to the states, we used the ChromHMM package to compute the overlap and neighbourhood enrichments of each state relative to coordinates of known gene annotations. The trained model was then used to compute the posterior probability of each state for each genomic bin in each condition. The regions were labelled using the state with the maximum posterior probability. The chromatin state models and browser tracks can be downloaded from http://www.broadinstitute.org/~anshul/projects/liz/segmentation/results/S14/webpage_14.html.

Differential analysis and visualization. We used the DESeq2 method that models read count statistics from replicates across multiple conditions to identify differentially expressed genes and regions of enrichment of histone marks³⁷. Our procedures are consistent with the standards for ChIP-seq and RNA-seq analysis determined by rigorous benchmarking as a part of the ENCODE project³³. The minimal recommended depth for sufficient sensitivity of peak detection for histone marks for the human or mouse genome is ~20 million mapped reads³³. However, owing to limited amount of starting material obtained from a single mouse, we obtained ~10 million unique mapped reads from each biological replicate. Directly, using read counts from the original replicates would result in significant loss of power to detect differential events. To improve sensitivity, for each histone mark in each condition, we pooled mapped reads from all replicates and created a pair of pseudo-replicates with equal number of reads (~30 million) by randomly subsampling (without replacement) from the pool. Reads were then extended to the predominant fragment length. Extended-read counts were computed within all regions in the master peak list of a histone mark for all pseudo-replicates in all conditions and the table of counts was used as input to DESeq2. The raw data are available online (NCBI GEO GSE65159).

For RNA-seq data, the numbers reads overlapping ENSEMBL gene models³⁸ were determined by HT-Seq (<http://www-huber.embl.de/users/anders/HTSeq/>). The raw data are available online (NCBI GEO GSE65159). To ensure that the genes we chose were sufficiently quantifiable, we remove every gene where fewer than 20 reads were found across all samples. The resulting set of genes is found in Supplementary Table 1.

IGV³⁹ is used to visualize the histone marks, gene expression, chromatin state and AD GWAS data relative to the RefSeq gene model. Gene expression levels shown are raw read density. Levels of histone marks plotted are the log-likelihood ratio of ChIP signal relative to whole-cell extract control.

Within the DESeq2 framework of generalized linear models, we used a combination of different models to determine the significantly regulated genes and significantly regulated histone mark levels. We compared the set of all 2-week and 6-week controls to the three following groups: (1) the 2-week CK-p25 samples; (2) the 6-week CK-p25 samples; (3) a group containing both the 2-week and 6-week samples. The first two tests identified changes that might be 2-week or 6-week specific. The third test identified changes that might be too subtle to detect at any one time point alone. In each case, the most basic equation (count \approx CKp25 status) was used, but for a subset of samples. A stringent threshold of $q < 0.01$ (Benjamini Hochberg) was used to determine significantly changing genes expression levels and histone mark levels. Next, to determine the temporal bias of genes expression levels and histone marks we built another model (count \approx time), which compared the 2-week and 6-week CK-p25 samples. Levels considered likely to change ($q < 0.5$) were categorized as transient (2-week bias) or late-stage (6-week bias). The results of the RNA-seq analysis are found in Supplementary Table 1, while the results of the histone mark analysis are in Supplementary Table 2.

For the histone modifications, we defined promoters using H3K4me3 peaks labelled with the promoter state annotation under any of the conditions (CK-p25 or control, and 2 or 6 weeks). We define enhancers based on peaks of H3K27ac labelled by the enhancer chromatin state. We define Polycomb-repressed regions based on H3K27me3 peaks labelled by the Polycomb-repressed chromatin state. Our definitions are consistent with known roles of these histone modifications⁴⁰. Defining the boundaries of the regulatory regions using the peaks of the relevant histone modifications, and not the chromatin states, maximizes our power to detect changes in histone mark levels.

Pathway and Gene Ontology analysis for the gene expression data were then generated through the use of DAVID^{41,42}. We present the most significant biological process gene ontology category result as well as a subset of non-redundant less significant categories that still pass our threshold significant ($q < 0.01$). For the regulatory regions, GREAT (with default parameters) was used to find the fold enrichment in the same Gene Ontology categories⁴³.

Statistical framework for comparing CK-p25 changing genes and regulatory regions to other data sets. A common theme throughout the analysis is the characterization of regulatory regions that change in the CK-p25 mouse model. The most stringent control for this characterization is genes or regions of the same type that do not change in CK-p25. Owing to the six categories of direction (increasing and decreasing) and temporal pattern (transient, consistent and late-stage), we chose a discrete statistical framework as opposed to trying to define a ranking across these different conditions. To measure the overlap between these discrete categories and other discrete data sets, we could use either a hypergeometric P value or a binomial P value. For every test in the material described below, we computed both significance values and obtained consistent results, with only minor differences in exact P value. In general, we chose the hypergeometric test, which is the most direct to look at overlap of annotated regions. As opposed to the overlap of the CK-p25 mouse categories with other ChIP-Seq peaks, the overlap with transcription factor binding site motifs or SNPs can be thought of as sampling with replacement, which lends itself to the binomial P value. No power analysis was done to estimate sample size.

Comparison of histone marks and gene expression. As described above, DESeq2 was used to determine the log fold change in expression at 2 and 6 weeks in CK-p25 mice relative to control. Each enhancer and promoter was mapped to the closest ENSEMBL gene model based on distance to transcription start. For each category of histone mark direction and temporal pattern, we examined the enrichment of each category of CK-p25 gene expression change relative to unchanging genes. The significance of the enrichment is calculated using a hypergeometric test.

Identification of orthologous human regions. The promoter (H3K4me3 peaks annotated as transcription start site by chromatin state), enhancer (H3K27ac peaks annotated as enhancer by chromatin state) and Polycomb-repressed regions (H3K27me3 peaks annotated as Polycomb-repressed by chromatin state) were mapped to the human genome. BED files representing the coordinates of these peaks in mm9 were mapped to mm10 using liftOver⁴⁴. Those peaks were mapped compared to the human genome the UCSC multiple alignment chain files (<http://hgdownload.soe.ucsc.edu/goldenPath/mm10/multiz60way/>)⁴⁵. More specifically, the alignments that overlap the mouse peak and include hg19 were extracted. We calculated the human mouse pairwise alignment for each multiple alignment using the 'globalms' function of biopython (<http://biopython.org/>, version 1.59; python version 2.7.1). The highest scoring pairwise alignment formed base of the orthologous region in human. This region was extended on either side using lower scoring multiple alignments. The orthologous region in hg19 was required to be greater than 30 bp and no more than twice the length of the region in mouse. The mean conservation was examined using the PHASTCons score across placental mammals⁴⁶ based on the same 60-way multiple sequence alignment. The mapped enhancer regions were annotated with their chromatin state in human hippocampus, and across all 127 cell types and tissues, using BEDTools⁴⁷. The information from human tissues was collected according to protocols described in more detail in the companion publication as a part of the Roadmap Epigenomics project¹⁰ (<http://www.roadmapepigenomics.org/>). The protocols are approved by the NIH and no sequence information from identifiable subjects is provided.

Computational analysis of cell type proportion. To estimate computationally the relative composition of the neural and immune cell types we compared the changing expression patterns in our data set to a set of established cell-type-specific markers^{48–50}. This analysis shows that indeed it is likely that cell type composition is changing in the CK-p25 mouse model, consistent with a known decrease in number of neurons and astrogliosis at 6 weeks⁵. In summary, a transient enrichment of monocyte specific transcripts was observed at 2 weeks, a consistent enrichment of microglial-specific transcripts was enriched at 2 and 6 weeks, while astrocyte, oligodendrocyte and endothelial-specific markers were primarily increased at 6 weeks (Extended Data Fig. 9a, b). We could also detect a signature of neuronal loss, primarily at 6 weeks as well (Extended Data Fig. 9a, b). On the basis of these results alone, it is possible that changes in cell type composition are contributing to some of the differences we observe in our mouse model.

We also compared our data to a published study of microglial activation in another mouse model of AD⁸, to dissect out computationally changes that are probably due to cell type proportion versus changes due to activation within cells. If the changes in our mouse model were primarily due to cell type proportion, then the increase we observed in the CK-p25 mice should be proportional to the expression level of those genes microglia. If the changes we observed were primarily due to activation, then the changes we observe in the CK-p25 mouse should be proportional to the amount of activation found in during neurodegeneration⁸. Using the genes with published gene expression changes during activation⁸, we modelled these two possibilities as a linear regression problem and examined the relative significance of both hypotheses in the R programming language: CK-p25 log fold change \approx microglial expression + microglial activation log fold change. We found that the changes in the CK-p25 mice were significantly related to the changes in cell activation ($P = 2.9 \times 10^{-6}$) as

well as the changes in cell type proportion ($P = 2.7 \times 10^{-4}$), suggesting that both cell activation and composition changes occur.

Comparison of gene expression in mouse model and human AD. To examine the relationship between AD in the mouse model and human, we mapped each 1–1 orthologous gene from mouse to human in ENSEMBL (<http://www.biomart.org/>)⁵¹. For each category of expression change in mouse, we examined how that set of genes behaved in human AD cases relative to controls in whole hippocampus⁵² as well as laser capture microdissected hippocampal grey matter². To make this comparison we first downloaded both data sets from GEO (GSE1297 and GSE28146), applied a variance stabilization normalization, and then used limma⁵³ to find the log fold change in expression of all cases relative to controls. For each category of mouse gene expression, we calculated a P value based on a t -test for the bias of genes to increase or decrease in human AD relative to control. Because the original study⁵² had more confounders owing to changes in grey/white matter proportion, we focused our analysis on the 22 cases and 9 controls from the laser capture samples².

Enrichment of cofactors and transcription factors. Peaks representing both neural^{15,16} and immune^{17–19} enhancers or transcription factor binding were used to annotate the H3K27ac enhancers and H3K4me3 promoters. We used a hypergeometric test to evaluate whether or not these external annotations were enriched in the set of increased-level or decreased-level enhancers relative to the enhancers whose levels do not change. This same procedure was used to look at the enrichment of the CK-p25 enhancer orthologues in Roadmap Epigenome data. In this case, only enhancers that map to human are taken to be the background.

The putative binding sites based on transcription factor binding site motifs were identified independent of conservation and have been previously published⁵⁴. The transcription factor binding sites were further clustered based on similarity⁵⁵. The least significant of two statistical tests was used as a stringent measure of binding site enrichment. (1) The real transcription factor binding site motifs in the category of interest were compared shuffled control motifs that preserved nucleotide content. (2) The real transcription factor binding site motifs in the category of interest were compared the real motifs in enhancers that are stable in the CK-p25 mice. To estimate the significance for test (1), we use a binomial P value because the length distribution is different for changing regulator regions compared to unchanging. Then we estimate the probability of finding a site per base pair. To estimate the significance for test (2), we use a hypergeometric test. After identifying significant transcription factor binding sites in categories or regulatory regions, we collapsed the results into clusters of almost identical motifs, representing families. The group members can be found in a companion manuscript¹⁰ as well as online (<http://www.broadinstitute.org/~pouyak/motifs-table/>).

Luciferase reporter assay. A total of 14 oligonucleotide gBlocks (IDT), ranging in 500–1,000 nucleotides in length, and corresponding to 10 enhancer regions were synthesized. Each gBlock contained a constant 5'-GCTAGCCTCGAGGAT and 3'-ATCAAGATCTGGCCT region, for direct cloning into an EcoRV (NEB) linearized minimal promoter firefly luciferase vector pGL4.23[luc2/minP] (Promega). The resulting reporter constructs were verified by DNA sequencing. BV-2 cells were provided by B. Yankner. N2a cells were purchased from the American Type Culture Collection and maintained following their protocols. In brief, cells were grown in RPMI-1640 and DMEM respectively, supplemented with 10% FBS and 1% penicillin/streptomycin, and split 1:10 every 3 days. Cells were seeded into 24-well plates 1 day before transfection. Transfections into BV-2 and N2a cells were performed with 1 μ g of a pGL4.23 plasmid and 200 ng of Renilla luciferase construct pGL4.74[RLuc/TK] (Promega). Luciferase activities were measured 24 h after transfection using the Dual-Glo Luciferase Assay (Promega) and an EnVision 2103 Multilabel Plate Reader (PerkinElmer) and normalized to Renilla luciferase activity. All assays were performed in triplicate.

Microglia isolation. The 2-week-induced CK-p25 mice and age-matched controls were perfused with 50 ml PBS to wash away blood and minimize macrophage contamination in the brains. Hippocampal tissue was collected immediately after perfusion and a single-cell suspension was prepared as described previously⁵⁶. FACS was then used to purify CD11b⁺ CD45^{low} microglia cells using allophycocyanin (APC)-conjugated CD11b mouse clone M1/70.15.11.5 (Miltenyi Biotec, 130-098-088) and phycoerythrin (PE)-conjugated CD45 antibody (BD Pharmingen, 553081). Cells were collected directly into RNA lysis buffer (Qiagen, 74104).

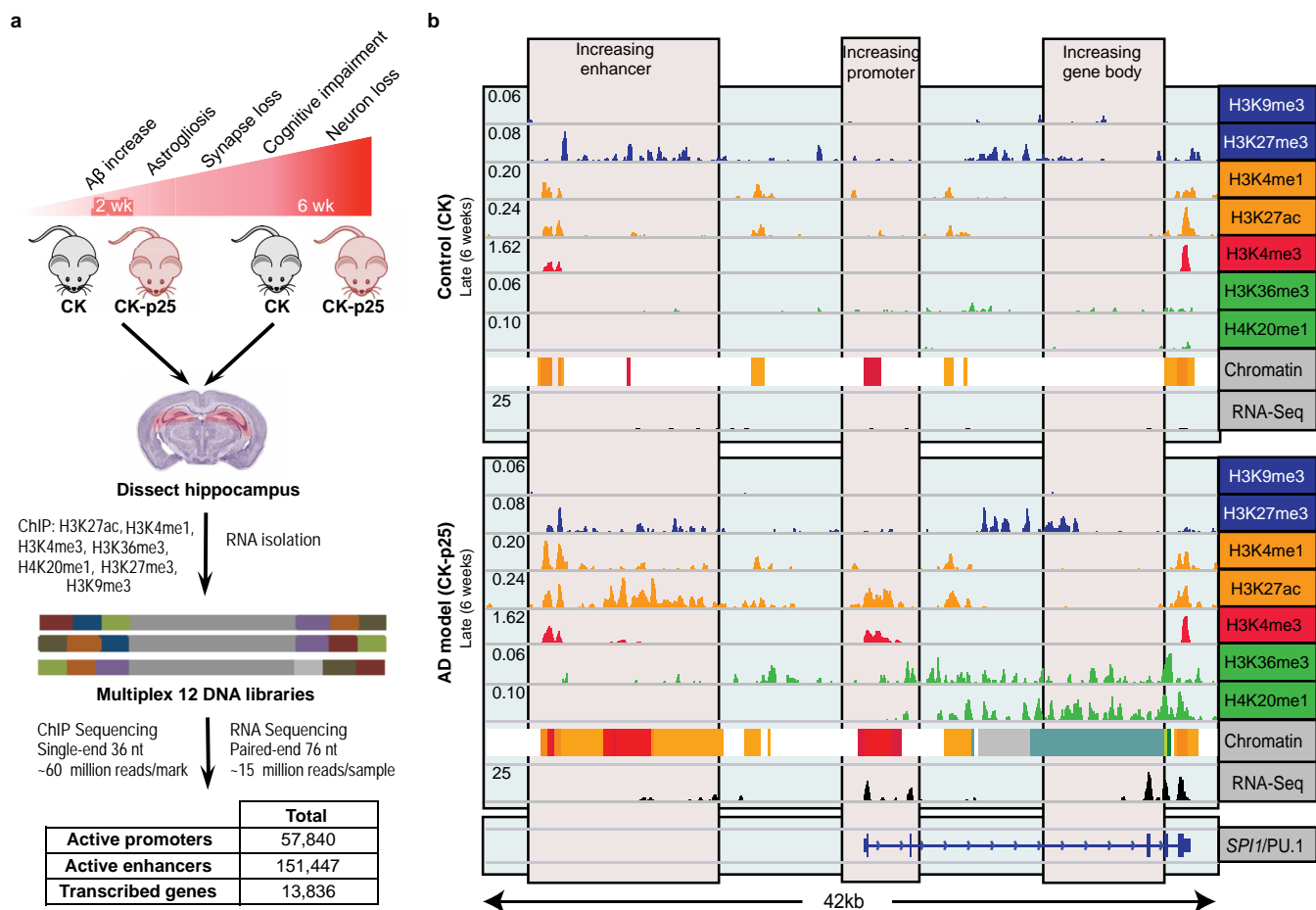
cDNA synthesis and qPCR. Total RNA was extracted using the RNeasy Mini kit (Qiagen, 74104) according to manufacturer's instructions. RNA concentration and purity was determined using Agilent's Bioanalyzer and reverse transcribed using iScript cDNA Synthesis Kit (Biorad, 170-8891). For gene expression analysis cDNA from three biological replicates was quantitatively amplified on a thermal cycler (BioRad) using SYBR green (Biorad) and gene-specific primers (Supplementary Table 8). The comparative C_t method⁵⁷ was used to examine differences in gene expression. Values were normalized to expression levels of *Cd11b* (also known as *Itgam*). Three technical replicates were used for each gene.

eQTL analysis. The human orthologous regions to mouse enhancers that change in the CK-p25 mouse were compared to control for their enrichment to overlap regulatory SNPs from published eQTL studies in immune cell types under a variety of conditions^{25,26}. Because the eQTLs were processed separately, we applied our own threshold ($P < 1 \times 10^{-4}$). We then calculated enrichment of human orthologues of different categories CK-p25 enhancers relative to stable regions and used a binomial P value to estimate the significance.

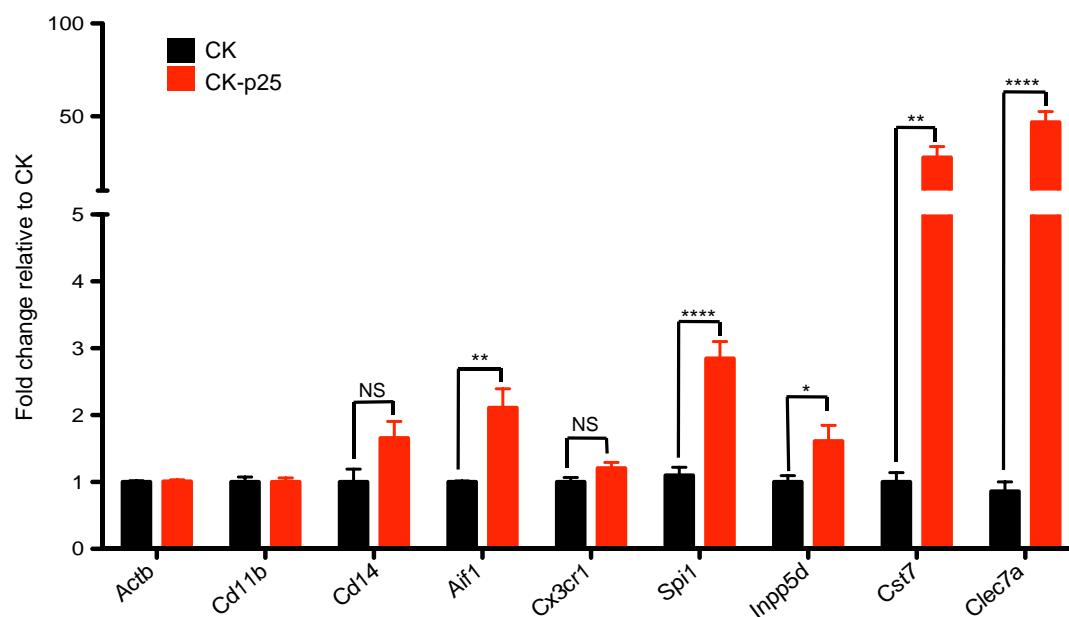
Enrichment of AD GWAS SNPs in Roadmap enhancers. The enrichment of AD GWAS SNPs that map to Roadmap enhancer regions is calculated on the basis of permutations of SNPs. In brief, SNPs were permuted 1,000,000 times preserving distance to gene, minor allele frequency, and a number of SNPs in LD. The thousand genomes projects database was used as the reference for this information.

Comparison of regulatory regions to AD meta-analysis. The enrichment of CK-p25 human enhancer orthologues in AD was calculated by comparing the number changing regions that overlap SNPs⁴ to unchanging regions that overlap SNPs. We calculate the significance using a binomial P value, in which the probability of success in the changing enhancers is based on the frequency in the unchanging enhancers. The results for the consistently increasing enhancers were slightly more significance when using a hypergeometric test instead of the binomial. To test whether the enrichment of increasing enhancer orthologous regions was due to the overlap with CD14⁺ cell enhancers, we repeated the above enrichment procedure within the set of CK-p25 enhancer orthologues that also overlap CD14⁺ cell enhancers. The enrichment using this control was still significant (3.0-fold enrichment, binomial $P = 1.3 \times 10^{-5}$). AD GWAS SNPs that were in a mouse enhancer orthologues were expanded using an LD of 0.8 and then tested for potential coding SNPs⁵⁸ or eQTLs (Supplementary Table 7).

31. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).
32. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18**, 1851–1858 (2008).
33. Landt, S. G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* **22**, 1813–1831 (2012).
34. Anders, S. *et al.* Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nature Protocols* **8**, 1765–1786 (2013).
35. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
36. Ernst, J., Kellis, M. & Chrom, H. M. M. Automating chromatin-state discovery and characterization. *Nature Methods* **9**, 215–216 (2012).
37. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).
38. Flicek, P. *et al.* Ensembl 2013. *Nucleic Acids Res.* **41**, D48–D55 (2013).
39. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).
40. Hoffman, M. M. *et al.* Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res.* **41**, 827–841 (2013).
41. Huang da, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* **4**, 44–57 (2009).
42. Huang da, W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37**, 1–13 (2009).
43. McLean, C. Y. *et al.* GREAT improves functional interpretation of *cis*-regulatory regions. *Nature Biotechnol.* **28**, 495–501 (2010).
44. Hinrichs, A. S. *et al.* The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.* **34**, D590–D598 (2006).
45. Blanchette, M. *et al.* Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14**, 708–715 (2004).
46. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
47. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
48. Zhang, Y. *et al.* An RNA-sequencing transcriptome and splicing database of glia, neurons, and vascular cells of the cerebral cortex. *J. Neurosci.* **34**, 11929–11947 (2014).
49. Hickman, S. E. *et al.* The microglial sensome revealed by direct RNA sequencing. *Nature Neurosci.* **16**, 1896–1905 (2013).
50. Butovsky, O. *et al.* Identification of a unique TGF- β -dependent molecular and functional signature in microglia. *Nature Neurosci.* **17**, 131–143 (2014).
51. Vilella, A. J. *et al.* EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.* **19**, 327–335 (2009).
52. Blalock, E. M. *et al.* Incipient Alzheimer's disease: microarray correlation analyses reveal major transcriptional and tumor suppressor responses. *Proc. Natl Acad. Sci. USA* **101**, 2173–2178 (2004).
53. Smyth, G. K., Michaud, J. & Scott, H. S. Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics* **21**, 2067–2075 (2005).
54. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476–482 (2011).
55. Kheradpour, P. & Kellis, M. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res.* **42**, 2976–2987 (2014).
56. Guez-Barber, D. *et al.* FACS purification of immunolabeled cell types from adult rat brain. *J. Neurosci. Methods* **203**, 10–18 (2012).
57. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta CT}$ method. *Methods* **25**, 402–408 (2001).
58. Ward, L. D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* **40**, D930–D934 (2012).

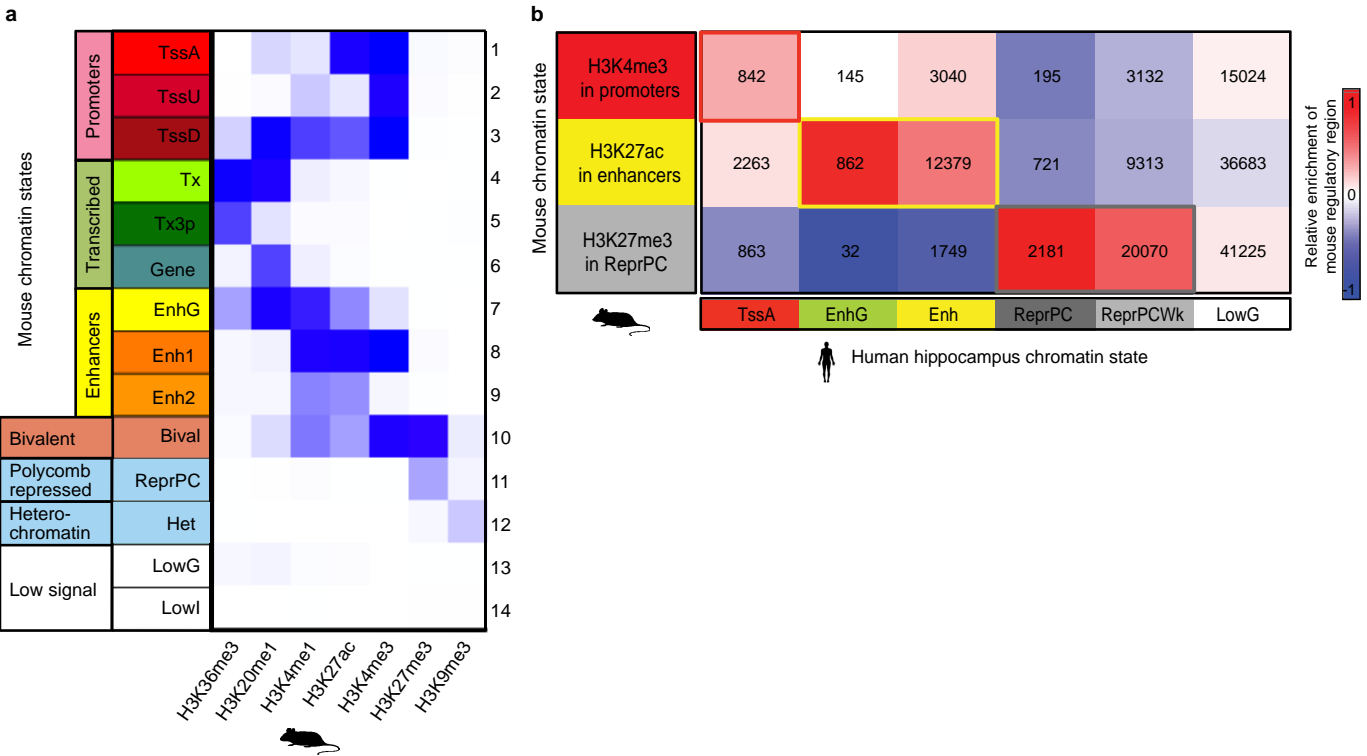


Extended Data Figure 1 | Epigenomic and transcriptomic profiling of a mouse model of AD. **a**, Experimental design and progression pathology in the CK-p25 mice. **b**, Gene expression and histone modification levels at the *SPI1* locus at 6 weeks of inducible p25 overexpression. Profiled are histone marks associated with repression (blue); histone marks associated with enhancers (orange); histone marks associated with promoters (red); histone marks associated with gene bodies (green); RNA-seq (black).



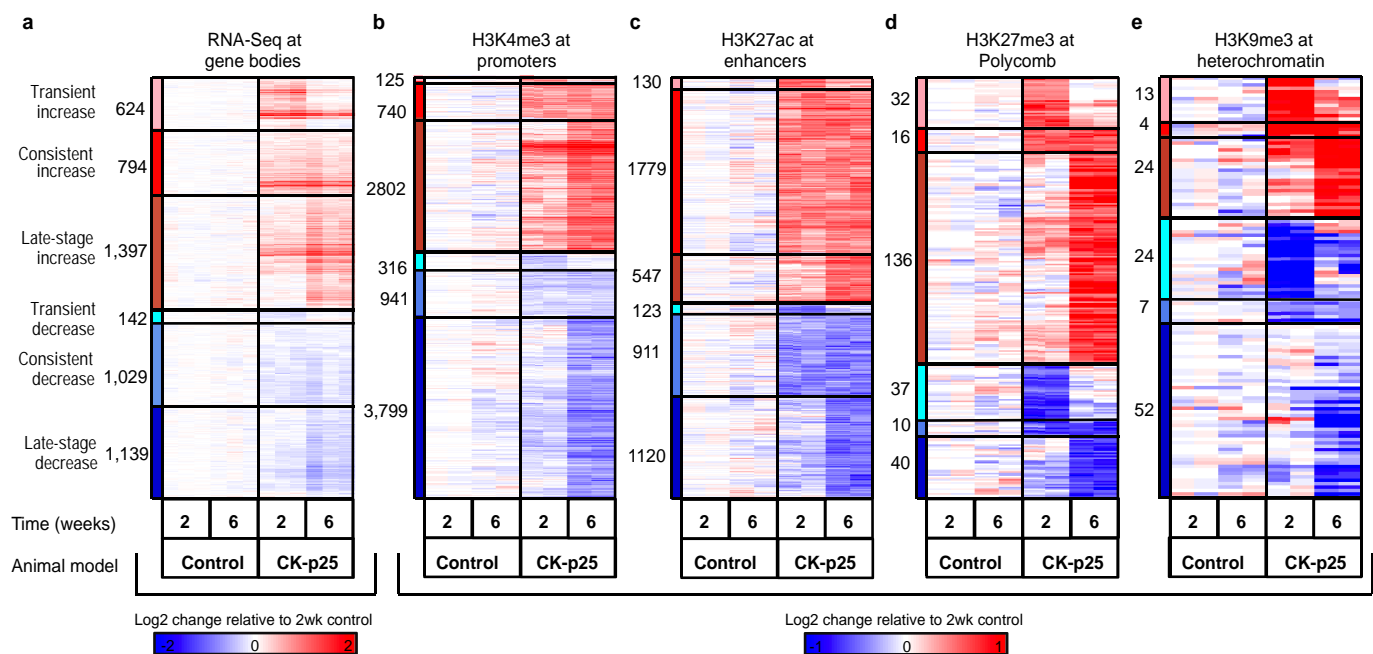
Extended Data Figure 2 | Differential microglia-specific gene expression changes in the CK-p25 mice. RT-qPCR of selected microglia markers and immune response genes shows upregulation of gene expression in fluorescence activated cell (FAC)-sorted CD11b⁺ CD45^{low} microglia from 2-week-induced

CK-p25 mice (red bars) relative to respective controls (black bars). *Actb* (β -actin) was used as a negative control. Values were normalized to *Cd11b* expression ($n = 3$, * $P < 0.05$, two-tailed t -test). NS, non-significant.



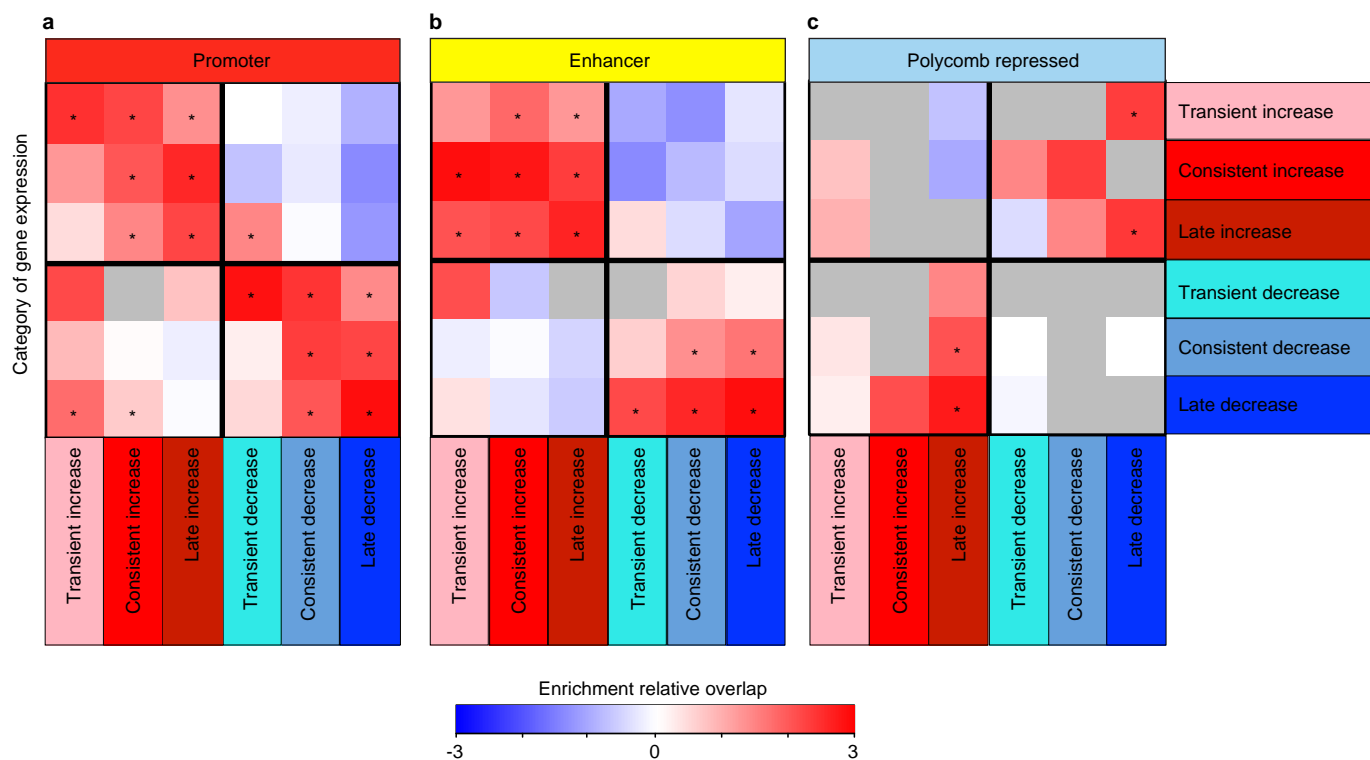
Extended Data Figure 3 | Chromatin state conservation. **a**, Combinatorial patterns of the seven histone modifications profiled were used to define promoter (1–3; A, active; D, downstream; U, upstream), gene body (4–6; tx, transcribed; 3P, 3 prime), enhancer (7–9; G, genic; 1 = strong, 2 = weak), bivalent (10), repressed Polycomb (11), heterochromatin (12), and low signal (13–14) chromatin states. Darker blue indicates a higher enrichment of the measured histone mark (*x* axis) to be found in a particular state (*y* axis).

b, Promoter, enhancer and repressed chromatin states in mouse hippocampus (rows), as profiled in this study, align to matching chromatin states in human (columns), as profiled by the Roadmap Epigenomics Consortium¹⁰. Shading indicates enrichment relative to human chromatin state abundance (columns). The number of regions overlapping is shown in each cell of the heatmap.



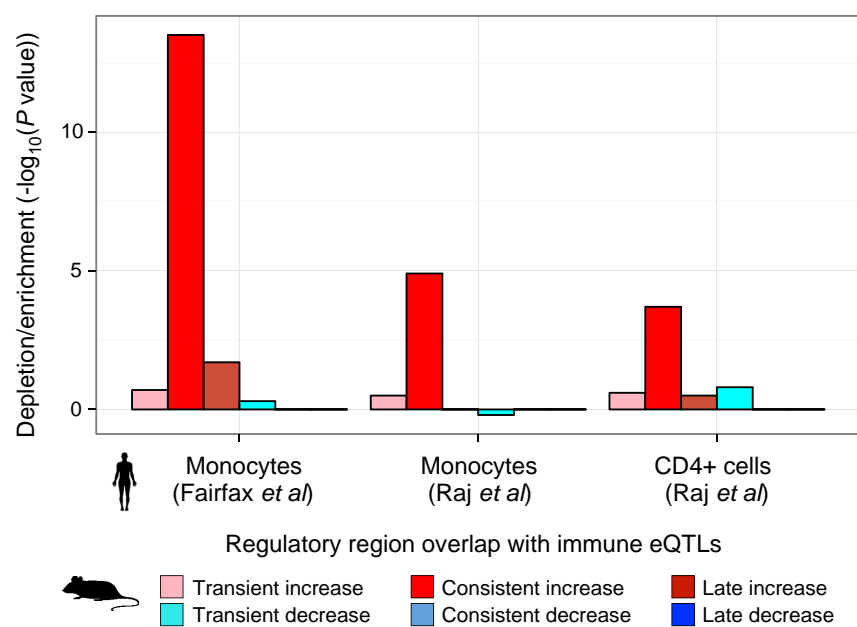
Extended Data Figure 4 | Differential gene expression and histone mark levels at regulatory regions in CK-p25 mice. a–e, Shown are six distinct classes of differentially modified regions: transient (early) increase (pink) or decrease (light blue), consistent increase (red) or decrease (blue), and late (6-week) increase (dark red) or decrease (navy blue). The heatmap shows the log fold change relative to 2-week controls for gene expression (a), H3K4me3

peaks at ‘TSS’ (transcription start site) chromatin state (b), H3K27ac peaks at enhancer chromatin state (c), H3K27me3 peaks overlapping the Polycomb repressed chromatin state (d), and H3K9me3 peaks overlapping the heterochromatin chromatin state (e). Numbers denote peaks falling into each category.

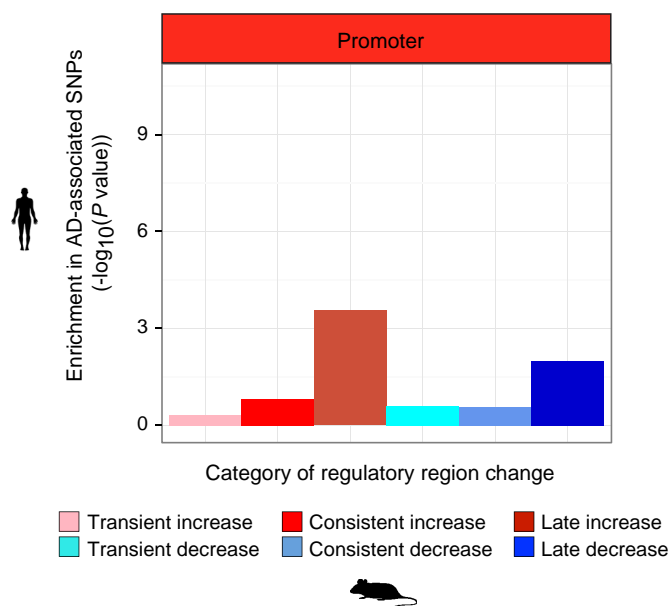


Extended Data Figure 5 | Relationship between changes of gene expression and regulatory regions in CK-p25 mice. a–c, For each class of gene expression change in the CK-p25 model (*x* axis), overlap with different histone modifications is shown (*y* axis) for H3K4me3 at promoters (a), H3K27ac at enhancers (b), and H3K27me3 at Polycomb repressed regions (c). Histone

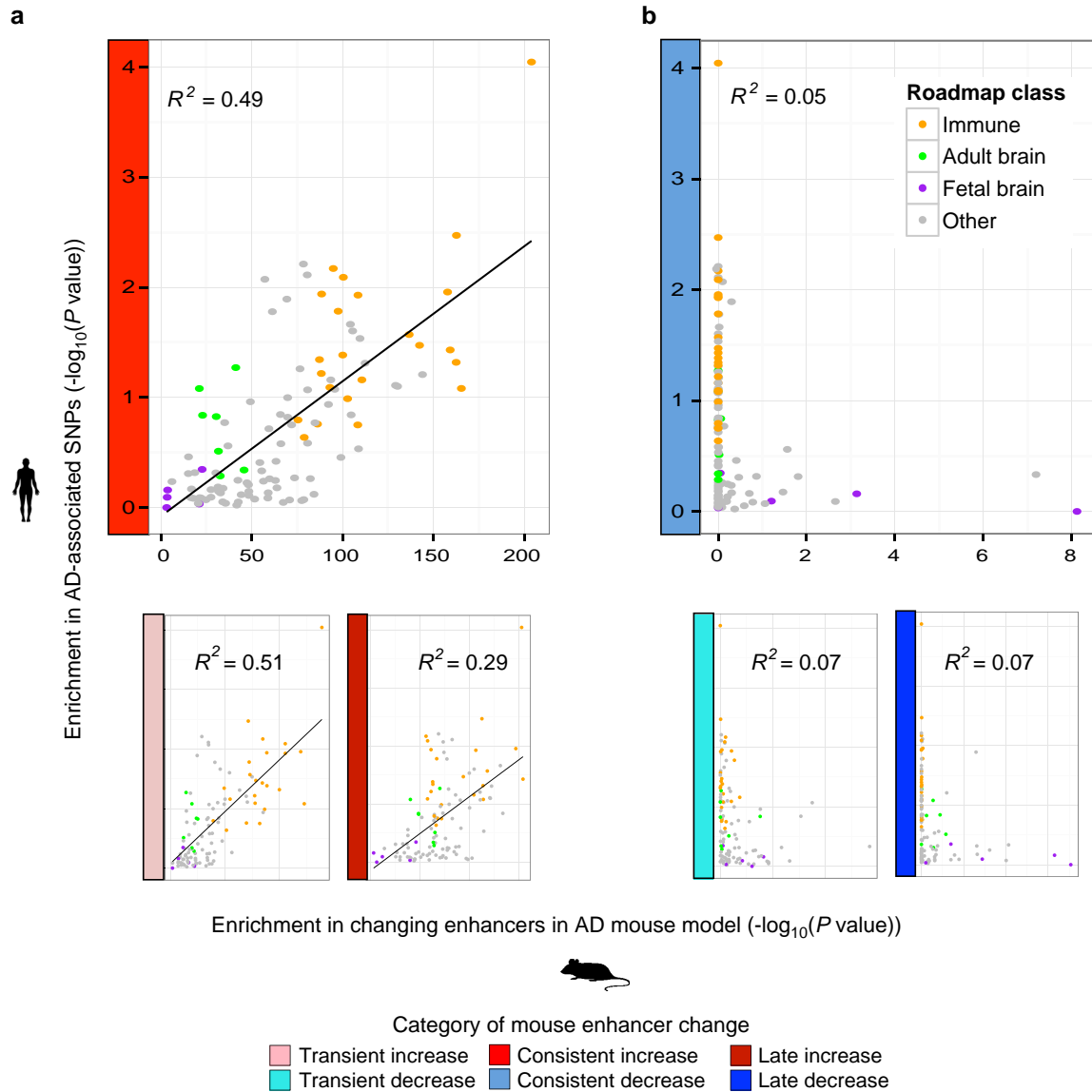
modifications were mapped to the nearest transcription start site (Supplementary Table 3) to show the enrichment of the changing regulatory regions relative to those that are stable in CK-p25. The significance is calculated based on the hypergeometric *P* value of the overlap.



Extended Data Figure 6 | Enrichment of immune cell eQTLs in increasing mouse enhancers. Enrichment of eQTL SNP (y axis; $-\log_{10}(\text{binomial } P < 10^{-4})$) in monocytes and CD4⁺ (refs 25, 26) is compared to the orthologous regions of CK-p25-affected enhancers relative to enhancers whose levels do not change.

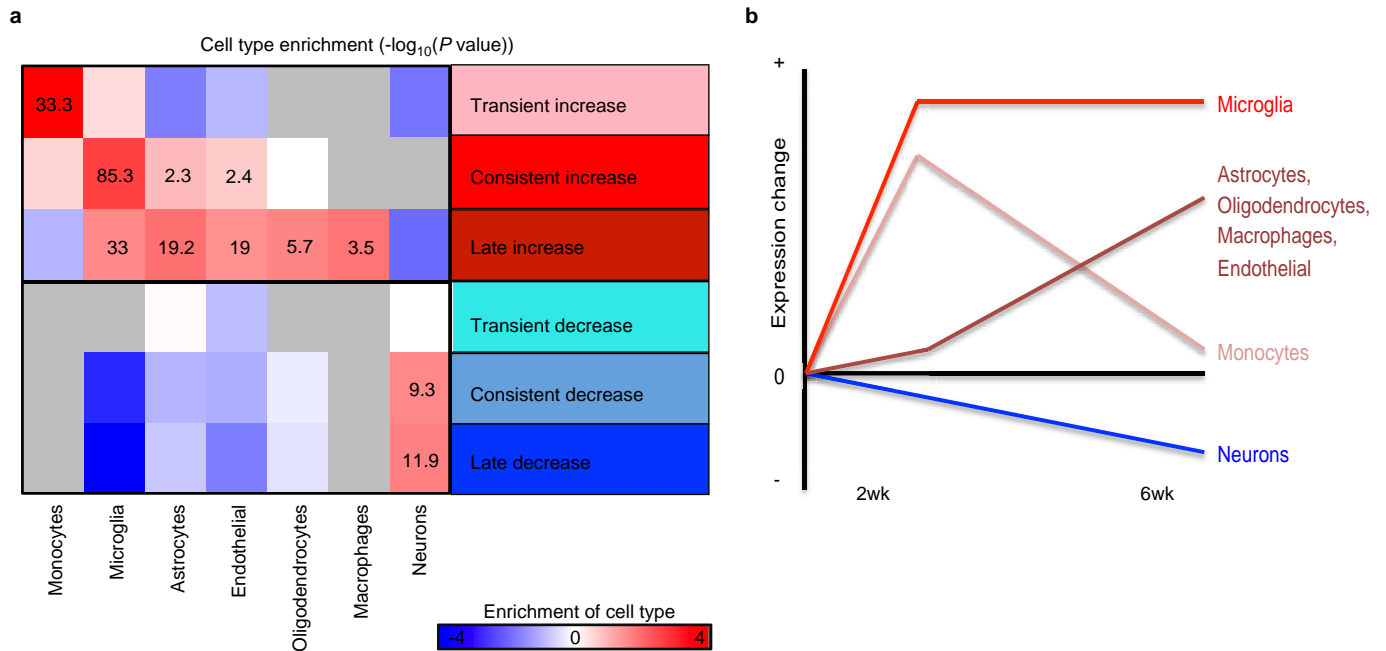


Extended Data Figure 7 | Weak enrichment of AD GWAS SNPs at differential CK-p25 promoters. Enrichment of AD-associated SNPs (y axis, binomial P value) in human regions orthologous to different classes of mouse promoters.



Extended Data Figure 8 | Enrichment of tissue-specific enhancer annotations from the Roadmap Epigenomics Consortium for AD-associated SNPs and mouse enhancers. a, b, Enrichment of AD-associated SNPs (y axis, permutation P value) in tissue-specific enhancer annotations from the Roadmap Epigenomics Consortium (points), relative to their

enrichment for increased-level (**a**) and decreased-level (**b**) (colours of different classes along y axis) of orthologous enhancer regions in the mouse AD model (x axis, hypergeometric P value). Linear regression trend line and R^2 based on Pearson correlation is shown.



Extended Data Figure 9 | Cell type composition. **a**, For each class of gene expression change (x axis), shown is the enrichment of cell-type-specific gene markers from published data sets^{48–50}. The macrophage and monocyte categories are computed relative to microglia^{49,50}. The enrichment is calculated

relative to the genes that do not change in expression level in the CK-p25 mice. Cells in the heatmap are labelled based on the $-\log_{10}(P \text{ value})$ (hypergeometric t -test). Cases where no genes overlapped are shown in grey. **b**, Summary of **a**, showing the inferred change in cell type composition across time.

Evolution of Darwin's finches and their beaks revealed by genome sequencing

Sangeet Lamichhane^{1*}, Jonas Berglund^{1*}, Markus Sällman Almén¹, Khurram Maqbool², Manfred Grabherr¹, Alvaro Martinez-Barrio¹, Marta Promerová¹, Carl-Johan Rubin¹, Chao Wang¹, Neda Zamani^{1,3}, B. Rosemary Grant⁴, Peter R. Grant⁴, Matthew T. Webster¹ & Leif Andersson^{1,2,5}

Darwin's finches, inhabiting the Galápagos archipelago and Cocos Island, constitute an iconic model for studies of speciation and adaptive evolution. Here we report the results of whole-genome re-sequencing of 120 individuals representing all of the Darwin's finch species and two close relatives. Phylogenetic analysis reveals important discrepancies with the phenotype-based taxonomy. We find extensive evidence for interspecific gene flow throughout the radiation. Hybridization has given rise to species of mixed ancestry. A 240 kilobase haplotype encompassing the *ALX1* gene that encodes a transcription factor affecting craniofacial development is strongly associated with beak shape diversity across Darwin's finch species as well as within the medium ground finch (*Geospiza fortis*), a species that has undergone rapid evolution of beak shape in response to environmental changes. The *ALX1* haplotype has contributed to diversification of beak shapes among the Darwin's finches and, thereby, to an expanded utilization of food resources.

Adaptive radiations are particularly informative for understanding the ecological and genetic basis of biodiversity^{1,2}. Those causes are best identified in young radiations, as they represent the early stages of diversification when phenotypic transitions between species are small and interpretable and extinctions are likely to be minimal³. Darwin's finches are a classic example of such a young adaptive radiation^{3,4}. They have diversified in beak sizes and shapes, feeding habits and diets in adapting to different food resources^{4,5} (Extended Data Table 1). The radiation is entirely intact, unlike most other radiations, none of the species having become extinct as a result of human activities⁴.

Fourteen of the currently recognized species evolved from a common ancestor in the Galápagos archipelago (Fig. 1a) in the past 1.5 million years according to mitochondrial DNA (mtDNA) dating⁶; a fifteenth species inhabits Cocos Island. The radiation proceeded rapidly as a result of strong isolation from the South American continent, generation of new islands by volcanic activity, climatic oscillations caused by the El Niño phenomenon, and sea level changes associated with glacial and interglacial cycles over the past million years that led to repeated alternations of island formation and coalescence^{7,8}.

Traditional taxonomy of Darwin's finches is based on morphology³, and has been largely supported by observations of breeding birds^{4,5} and genetic analysis^{6,9}. However, the branching order of several recently diverged taxa is unresolved⁶ and genetic analysis of phylogeny has been limited to mtDNA and a few microsatellite loci. Some candidate genes for beak development are differentially expressed in species with different beak morphologies^{10–12}, but the loci controlling genetic variation in beak diversity among Darwin's finches remain to be discovered.

Here we report results from whole genome re-sequencing of 120 individuals representing all Darwin's finch species and two closely related tanagers, *Tiaris bicolor* and *Loxigilla noctis*¹³. For some species we collected samples from multiple islands (Fig. 1a). We comprehensively analyse patterns of intra- and interspecific genome diversity and phylogenetic relationships among species. We find widespread evidence of interspecific gene flow that may have enhanced evolutionary

diversification throughout phylogeny, and report the discovery of a locus with a major effect on beak shape.

Considerable nucleotide diversity

We generated approximately 10× sequence coverage per individual bird using 2 × 100 base-pair (bp) paired-end reads (Extended Data Fig. 1). Reads were aligned to the genome assembly of a female medium ground finch (*G. fortis*)¹⁴. We identified Z- and W-linked scaffolds on the basis of significant differences in read depth between males (ZZ) and females (ZW) (Supplementary Table 1) and generated a *G. fortis* mtDNA sequence through a combined bioinformatics and experimental approach. Stringent variant calling revealed approximately 45 million variable sites within or between populations. We found a considerable amount of genetic diversity within each population, in the range 0.3×10^{-3} to 2.2×10^{-3} (Extended Data Table 2), similar to that reported in other bird populations¹⁵ including island populations of the zebra finch¹⁶. We used these estimates of diversity to estimate effective population sizes of Darwin's finch species within a range of 6,000–60,000 (Supplementary Text). Extensive sharing of genetic variation among populations was evident, particularly among ground and tree finches, with almost no fixed differences between species in each group (Extended Data Fig. 2).

Genome-based phylogeny

According to the classical taxonomy of Darwin's finches, supported by morphological and mitochondrial (cytochrome b) data, warbler finches were the first to branch off, and ground and tree finches constitute the most recent major split^{3,6,9}. Our maximum-likelihood phylogenetic tree based on autosomal genome sequences is generally consistent with current taxonomy, but shows several interesting deviations (Fig. 1b). First, *Geospiza difficilis* occurring on six different islands forms a polyphyletic group separated into three distinct groups: (1) populations occupying the highlands of Pinta, Santiago and Fernandina, (2) populations occupying the low islands of Wolf and Darwin in the northwest^{3,6,9} and (3) the population on Genovesa in the northeast. This is consistent with

¹Department of Medical Biochemistry and Microbiology, Uppsala University, SE-751 23 Uppsala, Sweden. ²Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, SE-75007 Uppsala, Sweden. ³Department of Plant Physiology, Umeå University, SE-901 87 Umeå, Sweden. ⁴Department of Ecology and Evolutionary Biology, Princeton University, Princeton, New Jersey 08544, USA. ⁵Department of Veterinary Integrative Biosciences, Texas A&M University, College Station, Texas 77843-4458, USA.

*These authors contributed equally to this work.

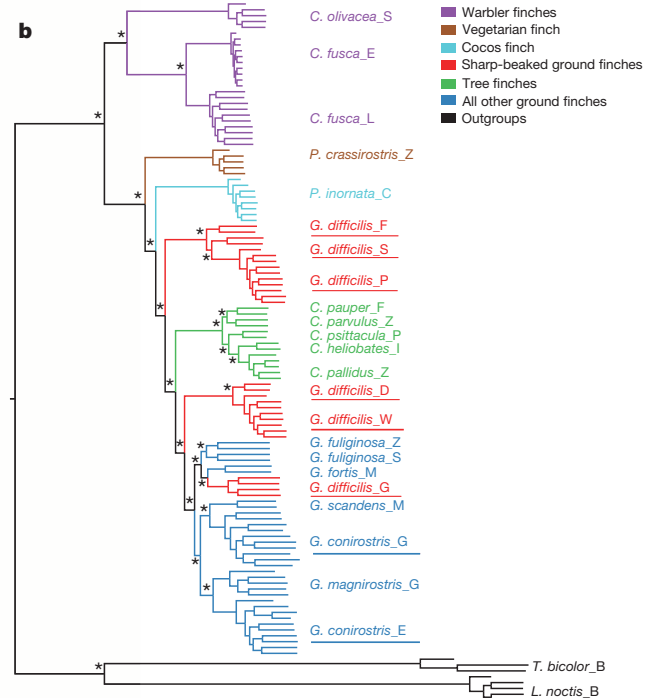
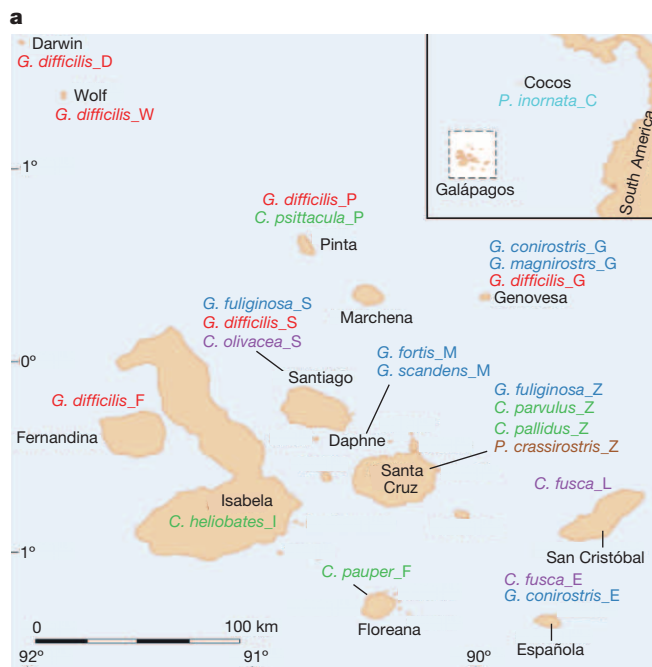


Figure 1 | Sample locations and phylogeny of Darwin's finches.

a, Geographical origin of samples; the letter after the species name is the abbreviation used for geographical origin. The map is modified from ref. 30. **b**, Maximum-likelihood trees based on all autosomal sites; all nodes having full

local support on the basis of the Shimodaira–Hasegawa test are marked by asterisks. The colour code for groups of species applies to both panels. Taxa that showed deviations from classical taxonomy are underscored.

an earlier version of the taxonomy, in which these three groups were classified as distinct species on the basis of morphological differences^{17,18}.

Second, *Geospiza conirostris* on Española showed the highest genetic similarity to another species, *Geospiza magnirostris*, whereas *G. conirostris* on Genovesa clustered with *Geospiza scandens* (Fig. 1b). Here, phenotypic similarity parallels genetic similarity; *G. conirostris* on Genovesa have a pointed beak similar to *G. scandens*, whereas those on Española have a blunt beak more similar to the beaks of *G. magnirostris* (Extended Data Fig. 3).

A network constructed from autosomal genome sequences indicates conflicting signals in the internal branches of ground and tree finches that may reflect incomplete lineage sorting and/or gene flow (Extended Data Fig. 3). The exact branching order of the most recently evolved ground and tree finches should be interpreted with caution as it may change with additional sampling. Since our data revealed some important discrepancies with the phenotype-based taxonomy, we propose a revised taxonomy for the sharp-beaked ground finch (*G. difficilis*) and the large cactus finch (*G. conirostris*) (Supplementary Text and Extended Data Fig. 4), but will use the current names in the text.

We dated phylogenetic splits on the basis of genome divergence (Fig. 2a), and compared these estimates with those obtained using mtDNA (Extended Data Fig. 5a and Supplementary Text). We infer that the most basal split, between warbler finches (*Certhidea* sp.) and other finches, occurred about 900,000 years ago. The rapid radiations of ground and tree finches began around 100,000–300,000 years ago. Although these estimates are based on whole-genome data, they should be considered minimum times, as they do not take into account gene flow.

Extensive interspecies gene flow

The discrepancies between phylogenies based on morphology and genome sequences may be due to convergent evolution and/or interspecies gene flow. We found evidence of introgression from three sources: ABBA–BABA tests, discrepancies between phylogenetic trees based on autosomal and sex-linked loci, and mtDNA (Supplementary Text and Extended Data Fig. 5a).

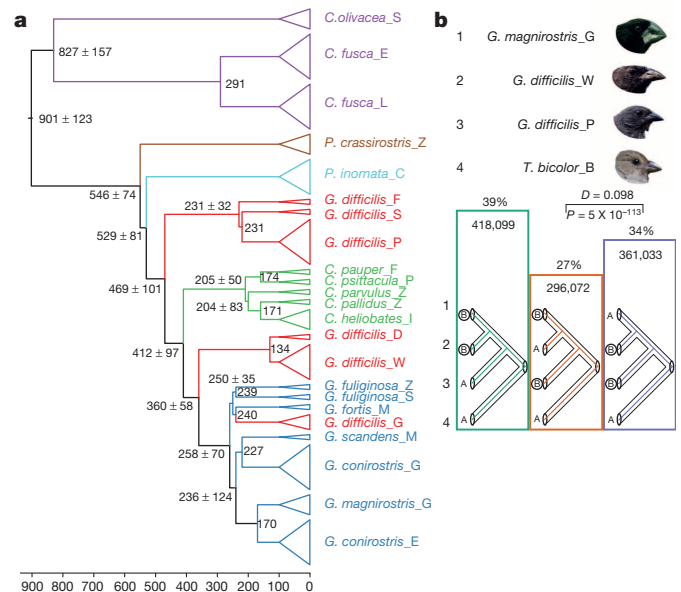


Figure 2 | Population history. **a**, Dating the nodes (in thousands of years) with confidence intervals (when applicable) in the phylogeny on the basis of divergence corrected for coalescence in ancestral populations; the topology is the representation of the inferred species tree from Fig. 1b. **b**, ABBA–BABA analysis of *G. magnirostris*, *G. difficilis* from Wolf and Pinta, and *L. noctis*. Number of sites supporting different trees is indicated both as a percentage and as actual numbers. The *D* statistic and corresponding Holm–Bonferroni-corrected *P* value are given for testing the null hypothesis of symmetry in genetic relationships. Finch heads are reproduced from ref. 5. *How and Why Species Multiply: The Radiation of Darwin's Finches* by Peter R. Grant & B. Rosemary Grant. Copyright © 2008 Princeton University Press. Reprinted by permission.

First, the D statistic¹⁹ associated with the ABBA–BABA test was used to compare two populations of *G. difficilis* from Pinta and Wolf, and *G. magnirostris* from Genovesa, using *L. noctis* as outgroup; *G. magnirostris* also occurs on Wolf but we lacked samples from that population. The analysis confirmed that *G. difficilis* on Wolf has a closer genetic relationship with *G. magnirostris* than with *G. difficilis* on Pinta (Fig. 2b). But there is evidence of gene flow between *G. difficilis* on Wolf and Pinta ($P = 5 \times 10^{-113}$), because the substantial asymmetry in genetic relationships cannot be explained by incomplete lineage sorting. However, the D statistic does not distinguish admixture from ancestral subdivision¹⁹. We conclude that the closely related populations of *G. difficilis* on Wolf and Darwin are a species of mixed ancestry where most of the genome originates from *G. magnirostris* or a close relative (Supplementary Table 2), whereas a considerable proportion of the genome, possibly including genetic variants affecting phenotypic characters, is derived from *G. difficilis*. Similarly, *G. difficilis* on Genovesa shows a closer genetic relationship to the other ground and tree finches than to *G. difficilis* on Pinta, but we also found evidence for gene flow between the two groups previously classified as *G. difficilis* ($P = 3 \times 10^{-87}$; Supplementary Table 2).

We next investigated gene flow involving the populations of *G. conirostris* on Genovesa and Española, which appear as separate species in our phylogenetic analysis. The ABBA–BABA analysis confirmed that *G. conirostris* on Española shows a closer genetic relationship to *G. magnirostris* than to *G. conirostris* on Genovesa (Extended Data Fig. 6a), but also provided evidence for gene flow between *G. conirostris* on Española and *G. conirostris* on Genovesa, which may explain some of their phenotypic similarities and their previous classification as a single species.

Given the evidence of relatively recent hybridization, we explored the possibility of more ancient hybridization between warbler finches (*Certhidea fusca* and *Certhidea olivacea*) and other finches. ABBA–BABA analysis provided evidence for gene flow between *C. fusca* and the other finches ($P = 7 \times 10^{-199}$; Extended Data Fig. 6b). This pattern of gene flow was apparent for all non-warbler finches, implying that it occurred before the radiation of the non-warbler finches (Supplementary Table 2).

The trees based on autosomal (Fig. 1b) and Z-linked sites (Extended Data Fig. 5b) are not completely congruent. The tree based on Z-linked polymorphisms indicated that *G. difficilis* present on the highlands of Pinta, Fernandina and Santiago is more closely related to *Platyspiza crassirostris* and emerged before the Cocos finch split off from the ground and tree finches, whereas the autosomal tree indicates a reversed order for the emergence of the two species. This discrepancy can potentially be explained by gene flow between *G. difficilis* and tree and ground finches after the Cocos finch became reproductively isolated from the finches on the Galápagos, which affected Z-linked and autosomal loci to different degrees. It is a common observation in closely related species that there is more interspecies sharing of sequence polymorphisms at autosomal loci than at sex-linked loci²⁰. This interpretation of the phylogenetic status of *G. difficilis* (highland group) is supported by the trees based on both mtDNA and W (Extended Data Fig. 5), which suggest that *G. difficilis* diverged from the ancestor of other ground and tree finches before the emergence of the Cocos finch.

Finally, our analysis of demographic history using the pairwise sequentially Markovian coalescent (PSMC) model²¹ was consistent with extensive interspecies gene flow among the ground finches, as they have maintained larger effective population sizes than the other species (Supplementary Text and Extended Data Fig. 6c, d).

A major locus controlling beak shape

The most striking morphological difference among Darwin's finches concerns beak shape (Extended Data Fig. 3). We performed a genome-wide scan on the basis of populations that are closely related but show different beak morphology: *G. magnirostris* and *G. conirostris* on Española have blunt beaks, whereas *G. conirostris* on Genovesa and *G. difficilis*

on Wolf have pointed beaks. We used non-overlapping 15-kilobase (kb) windows to identify regions with the highest fixation indices (F_{ST}) between groups. The F_{ST} distribution was Z-transformed (ZF_{ST}) and regions with striking ZF_{ST} values were identified (Fig. 3a). Among the 15 most significant regions, six harboured genes previously associated with craniofacial and/or beak development in mammals or birds including calmodulin (*CALM*)¹¹, goosecoid homeobox (*GSC*)²², retinol dehydrogenase 14 (*RDH14*)²³, ALX homeobox 1 (*ALX1*)^{24,25}, fibroblast growth factor 10 (*FGF10*)²⁶ and forkhead box C1 (*FOXC1*)²⁷. A previous study demonstrated differential expression of *CALM* between finches with different beak types¹¹. Two other studies reported differential expression of bone morphogenetic protein 4 (*BMP4*)^{10,12}, but we did not observe any elevated ZF_{ST} values in the vicinity of this locus, suggesting that differential expression is controlled by other loci.

The most striking finding was a 240-kb region with high ZF_{ST} values, including the window with the highest ZF_{ST} score (9.46) overall (Fig. 3a, b). The region overlaps part of *LRRIQ1* (leucine-rich repeats and IQ motif containing 1), the entire *ALX1* gene and about 130 kb downstream of *ALX1*. No previous report indicates that *LRRIQ1* has a role during development in vertebrates. By contrast, *ALX1* is an excellent candidate for variation in beak morphology. It encodes a paired-type homeodomain protein that plays a crucial role in development of structures derived from craniofacial mesenchyme, the first branchial arch and the limb bud²⁴, and on migration of cranial neural crest cells, highly relevant to beak development²⁵. Loss of *ALX1* in humans causes disruption of early craniofacial development²⁴.

All individuals in the blunt beak category were homozygous for a blunt beak-associated haplotype (denoted *B*), except one heterozygous *G. conirostris* individual from Española. Furthermore, except for one heterozygous bird from Genovesa, all 19 *G. difficilis* individuals not included in the F_{ST} scan were homozygous for a pointed beak haplotype (*P*), consistent with their phenotypic appearance (sharp-beaked ground finches). This is notable because genome-wide, *G. difficilis* on Wolf, Darwin and Genovesa are all more closely related to the blunt-beaked *G. magnirostris* than to the pointed-beaked *G. difficilis* from Pinta (Fig. 2b).

A phylogenetic tree based on this region revealed a deep divergence between the *B* and *P* haplotypes that must have occurred soon after the split between warbler finches and other Darwin's finches (Fig. 3c). Apart from the blunt-beaked *G. magnirostris* and *G. conirostris* on Española, all individuals except three were homozygous for *P* haplotypes, the remaining three being heterozygous. The two *G. fortis* from Daphne Major Island were both homozygous, but for different haplotypes (*BB* and *PP*; Fig. 3c). The short branch lengths among *B* haplotypes are consistent with a selective sweep. There were 335 fixed differences between the *B* and *P* haplotypes (Fig. 3d, upper panel), which we assigned as derived or ancestral on the basis of comparison with the outgroup sequence (*L. noctis*). Derived alleles on the *B* haplotype were aggregated in the vicinity of *ALX1*, including the downstream region (Fig. 3d, middle panel). Furthermore, 8 of these 335 fixed differences occurred at conserved sites, and the *B* haplotype carried the derived allele at seven of them (Fig. 3d, lower panel). Four derived alleles occurred at sites corresponding to transcription factor binding sites in the human genome²⁸. Two other changes constitute missense mutations (L112P and I208V) at *ALX1* amino-acid residues that are highly conserved among birds and mammals (Extended Data Fig. 7), and 'Sorting Intolerant From Tolerant' (SIFT)²⁹ analysis classified both as damaging (score 0.03 for both). The ratio of non-synonymous to synonymous substitutions between the *P* and *B* alleles is high ($2/1 = 2.00$) compared with the ratio observed between the ancestral *P* allele and orthologous zebra finch ($2/14 = 0.14$) and human ($21/122 = 0.17$) sequences, suggesting that one or both of these missense mutations are non-neutral.

That *ALX1* is polymorphic in *G. fortis* (Fig. 3c, d, upper panel) is particularly interesting, because field observations have shown there is considerable diversity in beak shape in this species^{5,30}. We genotyped an additional 62 *G. fortis* birds from Daphne Major Island for a diagnostic

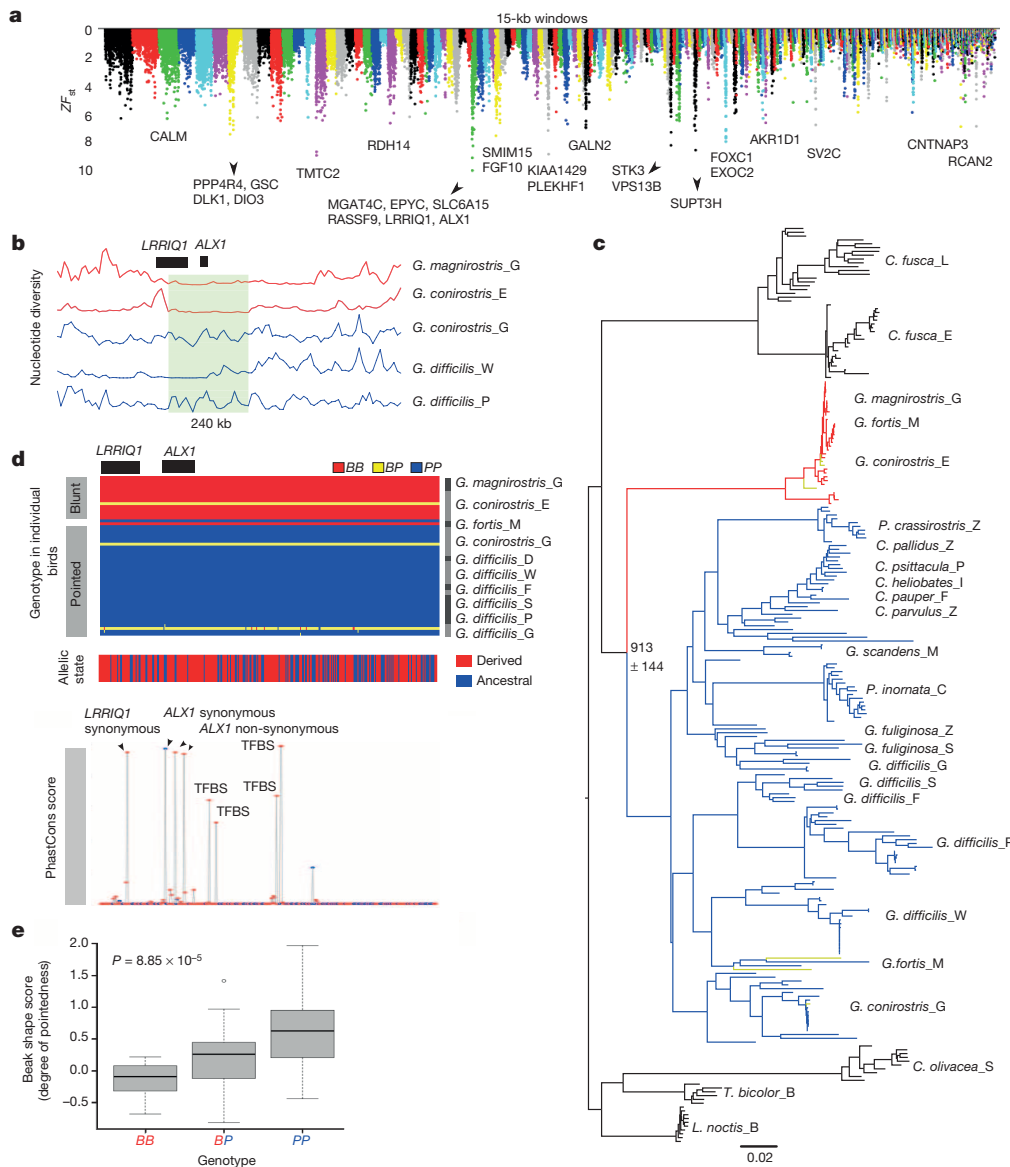


Figure 3 | A major locus controlling beak shape. **a**, Genome-wide F_{ST} screen comparing *G. magnirostris* and *G. conirostris* (Española) having blunt beaks with *G. conirostris* (Genovesa) and *G. difficilis* (Wolf) having pointed beaks. The y axis represents F_{ST} values. **b**, Nucleotide diversities in the *ALX1* region. The 240-kb region showing high homozygosity in blunt-beaked species is highlighted. Red and blue colours in **b–d** refer to blunt and pointed beak haplotypes, respectively. **c**, Neighbour-joining haplotype tree of *ALX1* region. Haplotypes originating from heterozygous birds (see text) are indicated in yellow. Estimated time since divergence (\pm confidence interval) of blunt and pointed beak haplotypes are given in thousands of years. **d**, Upper panel: genotypes at 335 SNPs showing complete fixation between *ALX1* haplotypes associated with blunt (B) and pointed (P) beaks. **d**, Middle panel: classification of alleles associated with blunt beaks at the 335 SNPs as derived or ancestral on the basis of allelic state in the outgroup. **d**, Lower panel: PhastCons³⁵ scores (on the basis of human, mouse and finch alignments) for the 335 SNP sites. TFBS, transcription factor binding sites. **e**, Linear regression analysis of beak-shape scores among *G. fortis* individuals on Daphne Major Island classified according to *ALX1* genotype; distribution of pointedness in each class is shown as a boxplot; $n = 62$; $F = 17.7$, adjusted $R^2 = 0.22$. Differences in six individual body and beak size traits were not significant (all $P > 0.05$).

single nucleotide polymorphism (SNP), and observed a significant association with beak shape ($P = 8.8 \times 10^{-5}$, Fig. 3e). *PP* homozygotes tended to have proportionately long, pointed beaks, *BB* homozygotes had proportionately deep, blunt beaks, whereas heterozygotes (*BP*) had intermediate beak shapes. We also compared haplotype frequencies among *G. fortis* individuals on Daphne Major Island with those on Santa Cruz, which have a larger and blunter beak on average³¹, possibly as a result of introgressive hybridization with *G. magnirostris*^{4,5}. We found the *B* haplotype to be more frequent on Santa Cruz than on Daphne Major (0.74, $n = 21$ versus 0.49, $n = 62$; $P = 0.007$, Fisher's exact test).

Natural selection on beak size and shape of *G. fortis* on Daphne Major Island has led to evolutionary change in the past few decades^{5,30}. Moreover, genetic variation in beak shape has been increased through introgressive hybridization^{5,30} with two species of *Geospiza*, *scandens* and *fuliginosa*, that have relatively pointed beaks. Therefore we expect hybrids and backcrosses in the *G. fortis* population to have a relatively high frequency of the *P* haplotype. We genotyped an additional 25 *G. fortis* at *ALX1*, added them to the sample of 62 (Methods) and compared the haplotype frequencies in eight hybrids (including backcrosses) and 79 non-hybrids. *ALX1-P* had a frequency of 0.75 among hybrids, and 0.44 among the others, which is statistically significant in the expected direction ($P = 0.03$, Fisher's exact test). Thus, *ALX1-P* alleles

introduced by introgressive hybridization most probably contributed to evolution of more pointed beaks in 1987 following natural selection as a result of a change in food supply in the 1985–86 drought³⁰.

Discussion

Our revised and dated phylogeny of Darwin's finches shows that the adaptive radiation took place in the past million years, with a rapid accumulation of species recently (Supplementary Text). We have genomically characterized the entire radiation, which has revealed a striking connection between past and present evolution. Evidence of introgressive hybridization, which has been documented as a contemporary process, is found throughout the radiation. Hybridization has given rise to species of mixed ancestry, in the past (this study) and the present³⁰. It has influenced the evolution of a key phenotypic trait: beak shape. Similar introgressive hybridization affecting an adaptive trait (mimicry) has been described in *Heliconius* butterflies³². The degree of continuity between historical and contemporary evolution is unexpected because introgressive hybridization plays no part in traditional accounts of adaptive radiations of animals^{1,2}. For young radiations it complements the better-known role of natural selection.

Charles Darwin first noted the diversity in beak shapes among the finches on Galápagos. Our genomic study has now revealed some of

the underlying genetic variation explaining this diversity. A polygenic basis for beak diversity is indicated by our discovery of about 15 regions with strong genetic differentiation between groups of finches with blunt or pointed beaks. We present evidence that the *ALX1* locus contributes to beak diversity, within and among species. The derived *ALX1-B* haplotype associated with blunt beaks has a long evolutionary history (hundreds of thousands of years), because its origin predates the radiation of vegetarian, tree and ground finches (Fig. 3c). This haplotype is fixed or nearly fixed in two ground finches with blunt beaks, *G. magnirostris* and *G. conirostris* on Española, and it co-segregates with variation in beak shape in *G. fortis*. As previously documented in domestic animals³³ and natural populations³⁴, the haplotype might have evolved by accumulating both coding and regulatory changes affecting *ALX1* function. Natural selection and introgression affecting this locus have contributed to the diversification of beak shapes among Darwin's finches and hence to their expanded utilization of food resources on Galápagos.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 9 October; accepted 31 December 2014.

Published online 11 February 2015.

- Schluter, D. *The Ecology of Adaptive Radiation* (Oxford Univ. Press, 2000).
- Seehausen, O. African cichlid fish: a model system in adaptive radiation research. *Proc. R. Soc. B* **273**, 1987–1998 (2006).
- Lack, D. *Darwin's Finches* (Cambridge Univ. Press, 1947).
- Grant, P. R. *Ecology and Evolution of Darwin's Finches* (Princeton Univ. Press, 1999).
- Grant, P. R. & Grant, B. R. *How and Why Species Multiply. The Radiation of Darwin's Finches* (Princeton Univ. Press, 2008).
- Petren, K., Grant, P. R., Grant, B. R. & Keller, L. F. Comparative landscape genetics and the adaptive radiation of Darwin's finches: the role of peripheral isolation. *Mol. Ecol.* **14**, 2943–2957 (2005).
- Ali, J. R. & Aitchison, J. C. Exploring the combined role of eustasy and oceanic island thermal subsidence in shaping biodiversity on the Galápagos. *J. Biogeogr.* **41**, 1227–1241 (2014).
- Geist, D., Snell, H., Snell, H., Goddard, C. & Kurz, M. in *The Galápagos: A Natural Laboratory for the Earth Sciences* (eds Harpp K. S., Mittelstaedt E., d'Ozouville N., & Graham, D.) 145–166 (American Geophysical Union, 2014).
- Farrington, H. L., Lawson, L. P., Clark, C. M. & Petren, K. The evolutionary history of Darwin's finches: speciation, gene flow, and introgression in a fragmented landscape. *Evolution* **68**, 2932–2944 (2014).
- Abzhanov, A., Protas, M., Grant, B. R., Grant, P. R. & Tabin, C. J. Bmp4 and morphological variation of beaks in Darwin's finches. *Science* **305**, 1462–1465 (2004).
- Abzhanov, A. *et al.* The calmodulin pathway and evolution of elongated beak morphology in Darwin's finches. *Nature* **442**, 563–567 (2006).
- Mallarino, R. *et al.* Two developmental modules establish 3D beak-shape variation in Darwin's finches. *Proc. Natl Acad. Sci. USA* **108**, 4057–4062 (2011).
- Burns, K. J. *et al.* Phylogenetics and diversification of tanagers (Passeriformes: Thraupidae), the largest radiation of Neotropical songbirds. *Mol. Phylogenet. Evol.* **75**, 41–77 (2014).
- Zhang, G., Parker, P., Li, B., Li, H. & Wang, J. The genome of Darwin's finch (*Geospiza fortis*). *GigaScience*, <http://dx.doi.org/10.5524/100040> (3 August 2012).
- Ellegren, H. The evolutionary genomics of birds. *Annu. Rev. Ecol. Evol. Syst.* **44**, 239–259 (2013).
- Balakrishnan, C. N. & Edwards, S. V. Nucleotide variation, linkage disequilibrium and founder-facilitated speciation in wild populations of the zebra finch (*Taeniopygia guttata*). *Genetics* **181**, 645–660 (2009).
- Swarth, H. S. The avifauna of the Galapagos Islands. *Occ. Pap. Calif. Acad. Sci.* **18**, 1–299 (1931).
- Lack, D. The Galapagos finches (Geospizinae): a study in variation. *Occ. Pap. Calif. Acad. Sci.* **21**, 1–159 (1945).
- Durand, E. Y., Patterson, N., Reich, D. & Slatkin, M. Testing for ancient admixture between closely related populations. *Mol. Biol. Evol.* **28**, 2239–2252 (2011).
- Qvarnstrom, A. & Bailey, R. I. Speciation through evolution of sex-linked genes. *Heredity* **102**, 4–15 (2009).
- Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011).
- Rivera-Perez, J. A., Wakamiya, M. & Behringer, R. R. Goosecoid acts cell autonomously in mesenchyme-derived tissues during craniofacial development. *Development* **126**, 3811–3821 (1999).
- Rowe, A., Richman, J. M. & Brickell, P. M. Retinoic acid treatment alters the distribution of retinoic acid receptor- β transcripts in the embryonic chick face. *Development* **111**, 1007–1016 (1991).
- Uz, E. *et al.* Disruption of *ALX1* causes extreme microphthalmia and severe facial clefting: expanding the spectrum of autosomal-recessive *ALX*-related frontonasal dysplasia. *Am. J. Hum. Genet.* **86**, 789–796 (2010).
- Dee, C. T., Szymoniuk, C. R., Mills, P. E. D. & Takahashi, T. Defective neural crest migration revealed by a zebrafish model of *Alx1*-related frontonasal dysplasia. *Hum. Mol. Genet.* **22**, 239–251 (2013).
- Brugmann, S. A. *et al.* Comparative gene expression analysis of avian embryonic facial structures reveals new candidates for human craniofacial disorders. *Hum. Mol. Genet.* **19**, 920–930 (2010).
- Sommer, P., Napier, H. R., Hogan, B. L. & Kidson, S. H. Identification of *Tgfb1i4* as a downstream target of *Foxc1*. *Dev. Growth Differ.* **48**, 297–308 (2006).
- Wang, J. *et al.* Factorbook.org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium. *Nucleic Acids Res.* **41**, D171–D176 (2013).
- Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols* **4**, 1073–1081 (2009).
- Grant, P. R. & Grant, B. R. *40 Years of Evolution. Darwin's Finches on Daphne Major Island* (Princeton Univ. Press, 2014).
- Boag, P. T. Growth and allometry of external morphology in Darwin's finches (*Geospiza*) on Isla Daphne Major, Galápagos. *J. Zool.* **204**, 413–441 (1984).
- The Heliconius Genome Consortium. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* **487**, 94–98 (2012).
- Andersson, L. Molecular consequences of animal breeding. *Curr. Opin. Genet. Dev.* **23**, 295–301 (2013).
- Linnen, C. R. *et al.* Adaptive evolution of multiple traits through multiple mutations at a single gene. *Science* **339**, 1312–1316 (2013).
- Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).

Supplementary Information is available in the online version of the paper.

Acknowledgements The National Science Foundation (USA) funded the collection of material under permits from the Galápagos and Costa Rica National Parks Services, and in accordance with protocols of Princeton University's Animal Welfare Committee. The map and images of finch heads are reproduced with permission from Princeton University Press. The project was supported by the Knut and Alice Wallenberg Foundation. Sequencing was performed by the SNP&SEQ Technology Platform, supported by Uppsala University and Hospital, SciLifeLab and Swedish Research Council (80576801 and 70374401). Computer resources were supplied by UPPMAX.

Author Contributions P.R.G. and B.R.G. collected the material. L.A., P.R.G. and B.R.G. conceived the study. L.A. and M.T.W. led the bioinformatic analysis of data. S.L. and J.B. performed the bioinformatic analysis with contributions from M.S.A., K.M., M.G., A.M.-B., C.-J.R. and N.Z. M.P. and C.W. performed experimental work. L.A., S.L., J.B., B.R.G., P.R.G. and M.T.W. wrote the paper with input from the other authors. All authors approved the manuscript before submission.

Author Information The Illumina reads have been submitted to the short reads archive (<http://www.ncbi.nlm.nih.gov/sra>) under accession number PRJNA263122 and the consensus sequence for the *G. fortis* mtDNA has been submitted to GenBank under accession number KM891730. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to L.A. (leif.andersson@imbim.uu.se).

METHODS

Study samples. No statistical methods were used to predetermine sample size. Blood samples from a total of 200 individuals of Darwin's finches, captured in mist nets and then released, were collected on FTA papers and stored at -70°C until DNA preparation. These included all 15 species of Darwin's finches currently present on the Galápagos and Cocos Island, and two closely related tanageres from Barbados used as outgroups¹³. Details on the name of each species, the specific island where they were sampled and the total number of individuals sampled from each species are in Extended Data Table 2 and phenotype descriptions of each species are in Extended Data Table 1.

Whole-genome sequencing. DNA was isolated from pieces of FTA papers using DNeasy tissue kit (QIAGEN). Each DNA sample was uniquely tagged with a sequence index during multiplexing library preparation protocol. The libraries (average fragment size about 400 bp) were sequenced using Illumina HiSeq2000 sequencers and 2×100 bp paired-end reads were generated. The amount of sequence per bird was targeted to approximately $10\times$ coverage.

Reference genome assembly. Sequence reads were aligned to the genome assembly of a female medium ground finch (*G. fortis*)¹⁴. This draft genome assembly has a size of ~ 1.07 Gb with scaffold N50 size of ~ 5.2 Mb and contig N50 size of ~ 30 kb. The annotation of the genome included a total of 16,286 protein-coding genes.

In addition, as the complete sequence for mtDNA was not previously available for any of the Darwin's finches, we also generated an assembly of the mtDNA genome sequence. For this, we first mapped all reads from one *G. fortis* individual against the zebra finch (*Taeniopygia guttata*) mtDNA. All the aligned reads were locally reassembled using SOAP DENOV³⁶, and then the gaps between the contigs were filled using Sanger sequencing to generate a single mtDNA genome sequence of 16.8 kb in length.

Sequence alignment and variant calling. The short sequence reads (2×100 bp) were quality checked using FASTQC (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>). Then we used BWA³⁷ (version 0.6.2) with default parameters to map the genomic reads from each individual against the reference genome assembly. The alignments were further checked for PCR duplicates using PICARD (<http://picard.sourceforge.net/>). We used Genome Analysis Toolkit (GATK)³⁸ for base quality recalibrations, insertion/deletion (INDEL) realignment, SNP and INDEL discovery and genotyping across all 120 samples simultaneously according to GATK best practice recommendations^{39,40}.

Quality filtering of the raw variant calls was done according to an in-house filtering pipeline that excluded a variant as low quality if it did not satisfy the following cut offs for filtering: SNP quality > 100 , base quality > 30 , mapping quality > 50 , haplotype score < 10 , Fisher strand bias < 60 , mapping quality rank sum > -4.0 , read position rank sum > -2.0 , quality by depth > 2.0 , minimum depth (summing all 120 samples) > 125 , and maximum depth (summing all 120 samples) $< 1,875$. These parameters are explained in detail in the GATK user manual³⁹. The cut-offs were chosen on the basis of the distribution of each of these parameters from the raw variant calls generated by the GATK UnifiedGenotyper module. The missing and low quality genotypes from the call set were inferred separately for each population using BEAGLE (version 3.3.2)⁴¹. Finally, we retained 44,753,624 variable sites in the data set. The variant calling in mtDNA was also performed using a similar BWA and GATK pipeline as described above. We identified 1,429 mtDNA variable sites in mtDNA. We calculated the average nucleotide diversity for autosomes, chromosomes Z and W, and in the mtDNA genome separately to estimate the amount of genetic variation in each population in different parts of the genome.

Identification of scaffolds from chromosomes Z and W. The medium ground finch genome assembly contains 27,239 scaffolds unassigned to chromosomes. We used the MultiSV package to identify scaffolds that belong to chromosomes Z and W by comparing the read depth for each scaffold in 85 males and 35 females. This analysis identified 133 scaffolds, which belonged to chromosome Z with a total length of 67,176,652 bp (Supplementary Table 1a), and 662 scaffolds, which belonged to chromosome W with a total length of 643,111 bp (Supplementary Table 1b).

Estimation of genetic distance and phylogeny reconstruction. We used PLINK (version 1.07)⁴² to calculate genetic distance (on the basis of proportion of alleles identical by state) for all pairs of individuals separately for autosomes and the Z chromosome. We used the neighbour-net method of SplitsTree4 (<http://www.splitstree.org/>) to compute the phylogenetic network from genetic distances. We used FastTree to infer approximately maximum-likelihood phylogenies with standard parameters for nucleotide alignments of variable positions in the data set (<http://meta.microbesonline.org/fasttree/>). FastTree computes local support values with the Shimodaira–Hasegawa test.

ABBA–BABA analysis. Patterns of gene flow and the extent of admixture in populations were analysed and tested for asymmetry in the frequencies of discordant gene trees in a three-population phylogeny rooted with an outgroup using the *D*

statistic⁴³ as implemented for polymorphic sites¹⁹. The *D* statistics were transformed to *Z* scores by division with the standard error, which was calculated with a jackknife procedure. Blocks of 40,000 variable sites for autosomes and 10,000 for the Z chromosome were used in the jackknife to overcome the effect of linkage disequilibrium, which yielded 1,027 and 291 blocks, respectively. The *Z* scores were translated to two-sided *P* values that were Holm–Bonferroni-corrected⁴⁴ for multiple testing by stepwise division of the lowest *P* value with the remaining number of tests performed for all 1,768 possible tests in the phylogeny and the two tests with pooled species (Supplementary Table 2).

Mutation rates. We used the following previously reported estimated mutation rates for nuclear and mtDNA: nuclear DNA, 2.04×10^{-9} per site per year estimated from the synonymous mutation rate on the Darwin's finches' lineage since the split from zebra finch⁴⁵; mtDNA, a fossil-calibrated divergence rate of 2.1% per million years for bird cytochrome b sequences⁴⁶.

Estimation of effective population size. Effective population sizes (N_e) were calculated from Watterson's θ (ref. 47) across the whole genome and the above-mentioned mutation rate. Fluctuations in N_e were inferred using PSMC³⁷ and with '64*1' as the time interval parameter pattern. Plots were scaled assuming a mutation rate per generation of 1.02×10^{-8} and a generation time of 5 years (ref. 48).

Dating the nodes in the phylogeny and demographic history. Times of population splits were calculated with our estimates of genetic distances in the two subtrees of a node and corrected for the time to coalescence in ancestral populations⁴⁹ and mutation rate. Confidence intervals were estimated from the standard deviation of genetic distances estimated from the pairwise species comparisons. We estimated the time of divergence between the blunt and pointed *ALX1* haplotypes by estimating the average pairwise difference at this locus between species containing all blunt and all pointed haplotypes and correcting for mutation rate. *G. fortis* and heterozygous individuals were excluded. Cytochrome b sequences were used to date the mtDNA phylogeny in which the most recently evolved ground finches (that is, *G. magnirostris*, *conirostris*, *scandens*, *fortis*, *fuliginosa* and *difficilis* on Genovesa) were treated as one population, with diversities averaged across species, because they did not form monophyletic groups according to species.

To elucidate and display the demographic history of Darwin's finches we used the pairwise sequentially Markovian coalescent (PSMC) model, which infers fluctuations in effective population size over evolutionary time from a single genome sequence²¹.

Signatures of selection for beak diversification. We scanned the whole genome in non-overlapping 15-kb windows to identify regions with increased genetic divergence (F_{ST}) between species with blunt and pointed beaks. We used VCFtools version 0.1.11 (ref. 50) to calculate F_{ST} . The genomic windows with high ZF_{ST} (> 6) were analysed for gene content.

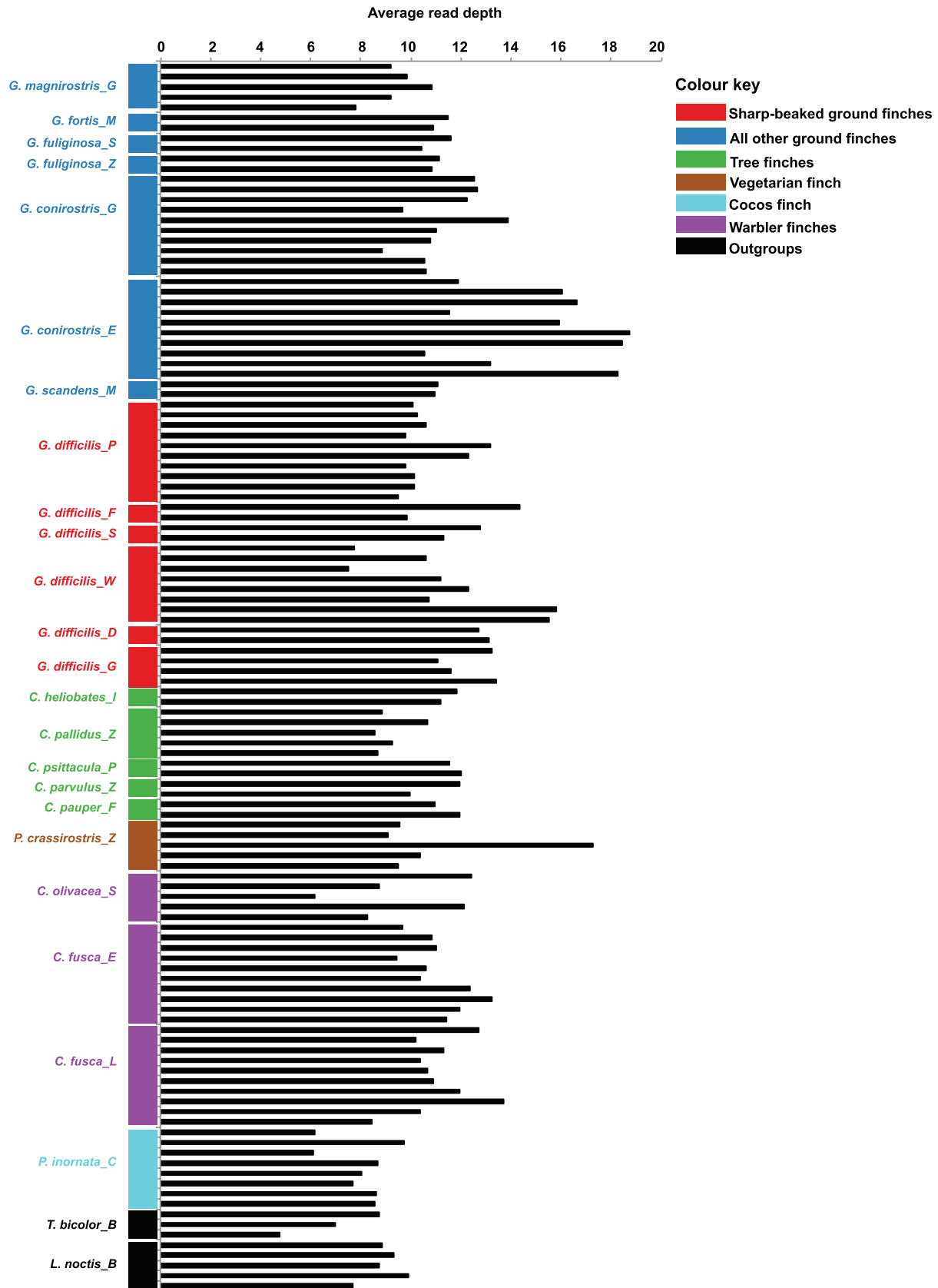
ALX1 genotyping in additional samples. A Taqman SNP genotyping assay (Life Technologies) was designed for one SNP (A/C at nucleotide position 517,149 bp in scaffold JH739921) diagnostic for the *ALX1* haplotypes associated with blunt and pointed beaks. A standard TaqMan Allele discrimination assay was performed using an Applied Biosystems 7900 HT real-time PCR instrument. The association of individual genotypes with beak shape measurements was evaluated using standard linear regression in R.

Comparison of ALX1 protein sequences among vertebrates. The *ALX1* protein sequence for *G. fortis* was downloaded from NCBI (XP_005421635). This *G. fortis* protein is a representative for the pointed allele and was edited to create a blunt counterpart by introducing the two amino-acid substitutions (L112P and I208V). *ALX1* protein sequences from other species were collected from predicted orthologues of the chicken *ALX1* gene in Ensembl⁵¹, including representative species from teleosts, reptiles, birds and mammals. The protein sequences were aligned using MUSCLE⁵² (version 3.8.31) with default settings, and the multiple sequence alignment was viewed and edited using Jalview^{29,53}. The probability of functional consequences of amino-acid substitutions was predicted using SIFT²⁹ with the multiple sequence alignment as input after exclusion of the blunt allele. Both substitutions were predicted to be damaging with probability scores of 0.03, where a score less than 0.05 is considered significant. Both predictions were reported to have a low confidence due to limited divergence in the alignment. However, we argue that because we have sampled orthologues from such a diverse set of species where *ALX1* displays considerable conservation, these predictions can be viewed with greater confidence. Protein domains were predicted with Interpro scan⁵⁴ using the *G. fortis* *ALX1* protein sequence.

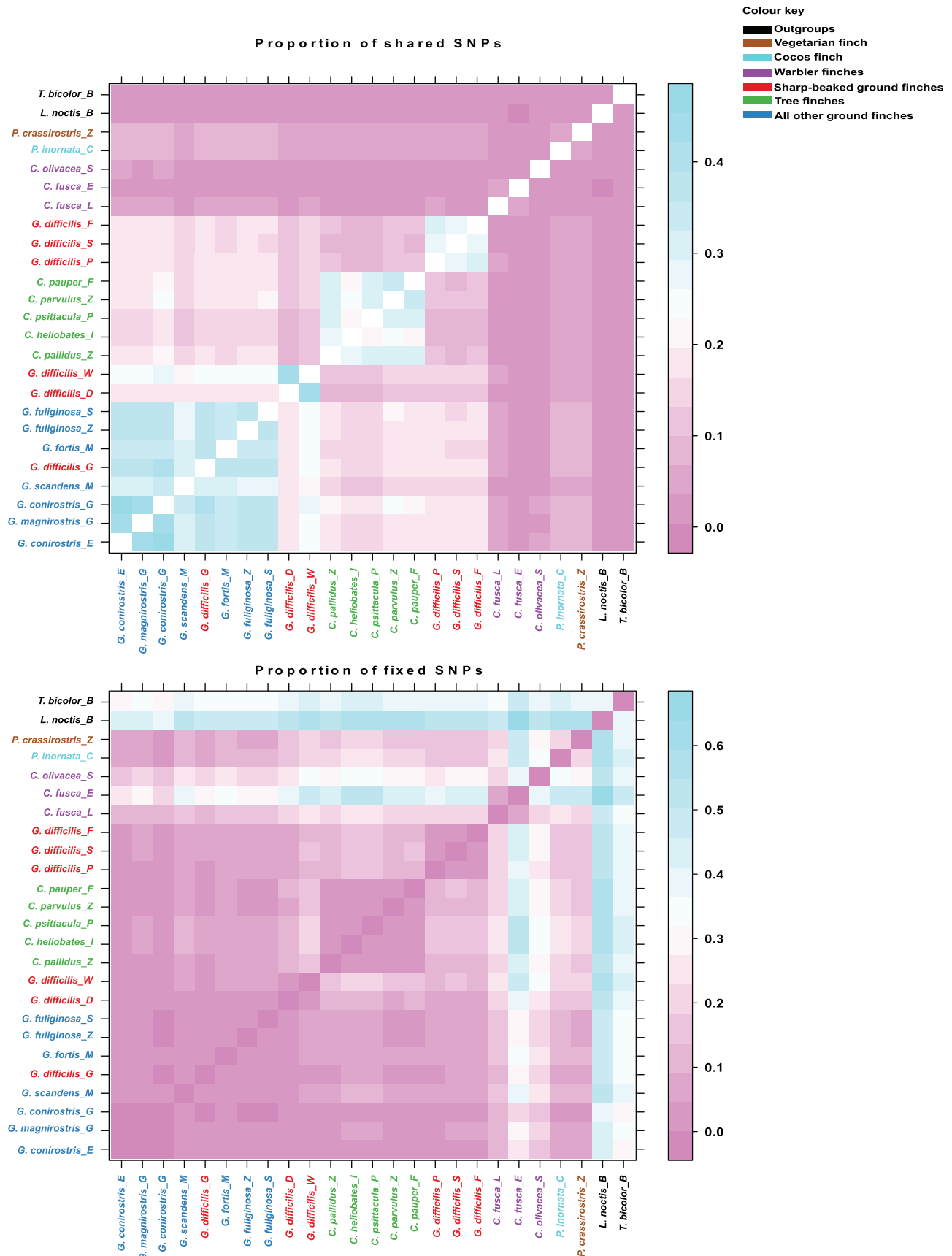
Functional annotation of SNPs. NCBI's genome annotation for the *G. fortis* assembly (GeoFor1) was downloaded from NCBI's FTP server (ftp://ftp.ncbi.nlm.nih.gov/genomes/Geospiza_fortis/) in GFF format. The annotation was filtered to include only genes annotated with a coding sequence (13,949 genes with 16,365 transcripts) before using it to build a local SnpEff (version 3.4) database⁵⁵. The SnpEff database was subsequently used to annotate all detected sequence variants among the Darwin's finches with putative functional effects according to categories

defined in the SnpEff manual. The upstream and downstream categories are regions within 5,000 bp in the respective direction of an annotated gene. SnpEff allows SNPs to be included in multiple categories; for example, a SNP may be intronic in one gene and a synonymous change in another gene residing in the intron of the first gene.

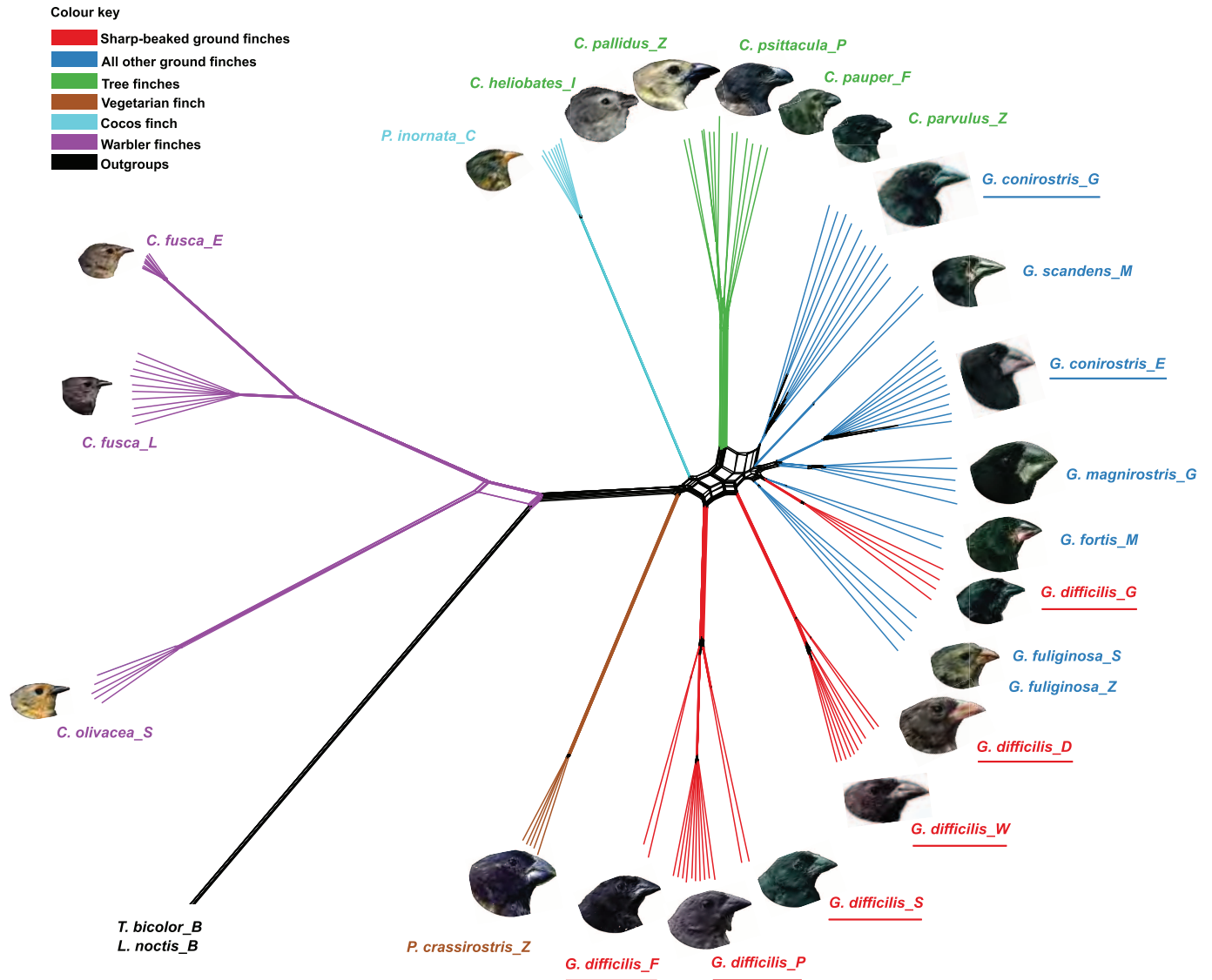
36. Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* **1**, 18 (2012).
37. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
38. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
39. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genet.* **43**, 491–498 (2011).
40. Van der Auwera, G. A. *et al.* From FastQ data to high-confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinform.* **43**, 11.10.1–11.10.33 (2002).
41. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**, 1084–1097 (2007).
42. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
43. Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010).
44. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **6**, 65–70 (1979).
45. Rands, C. *et al.* Insights into the evolution of Darwin's finches from comparative analysis of the *Geospiza magnirostris* genome sequence. *BMC Genomics* **14**, 95 (2013).
46. Weir, J. T. & Schluter, D. Calibrating the avian molecular clock. *Mol. Ecol.* **17**, 2321–2328 (2008).
47. Watterson, G. A. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**, 256–276 (1975).
48. Grant, B. R. & Grant, P. R. Demography and the genetically effective sizes of two populations of Darwin's finches. *Ecology* **73**, 766–784 (1992).
49. Nei, M. in *Molecular Evolutionary Genetics* 276–279 (Columbia Univ. Press, 1987).
50. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
51. Vilella, A. J. *et al.* EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.* **19**, 327–335 (2009).
52. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
53. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191 (2009).
54. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
55. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain *w¹¹¹⁸*; *iso-2*; *iso-3*. *Fly (Austin)* **6**, 80–92 (2012).
56. Grant, B. R., Grant, P. R. & Petren, K. The allopatric phase of speciation: the sharp-beaked ground finch (*Geospiza difficilis*) on the Galápagos islands. *Biol. J. Linn. Soc.* **69**, 287–317 (2000).
57. Grant, P. R., Abbott, I., Schluter, D., Curry, R. L. & Abbott, L. K. Variation in the size and shape of Darwin's finches. *Biol. J. Linn. Soc.* **25**, 1–39 (1985).
58. Schluter, D. & Grant, P. R. Ecological correlates of morphological evolution in a Darwin's finch, *Geospiza difficilis*. *Evolution* **38**, 856–869 (1984).
59. Rabosky, D. Diversity-dependence, ecological speciation, and the role of competition in macroevolution. *Ann. Rev. Evol. Ecol. Syst.* **44**, 481–502 (2013).



Extended Data Figure 1 | Read depth. Average read depth in all 120 samples of Darwin's finches and outgroup species.

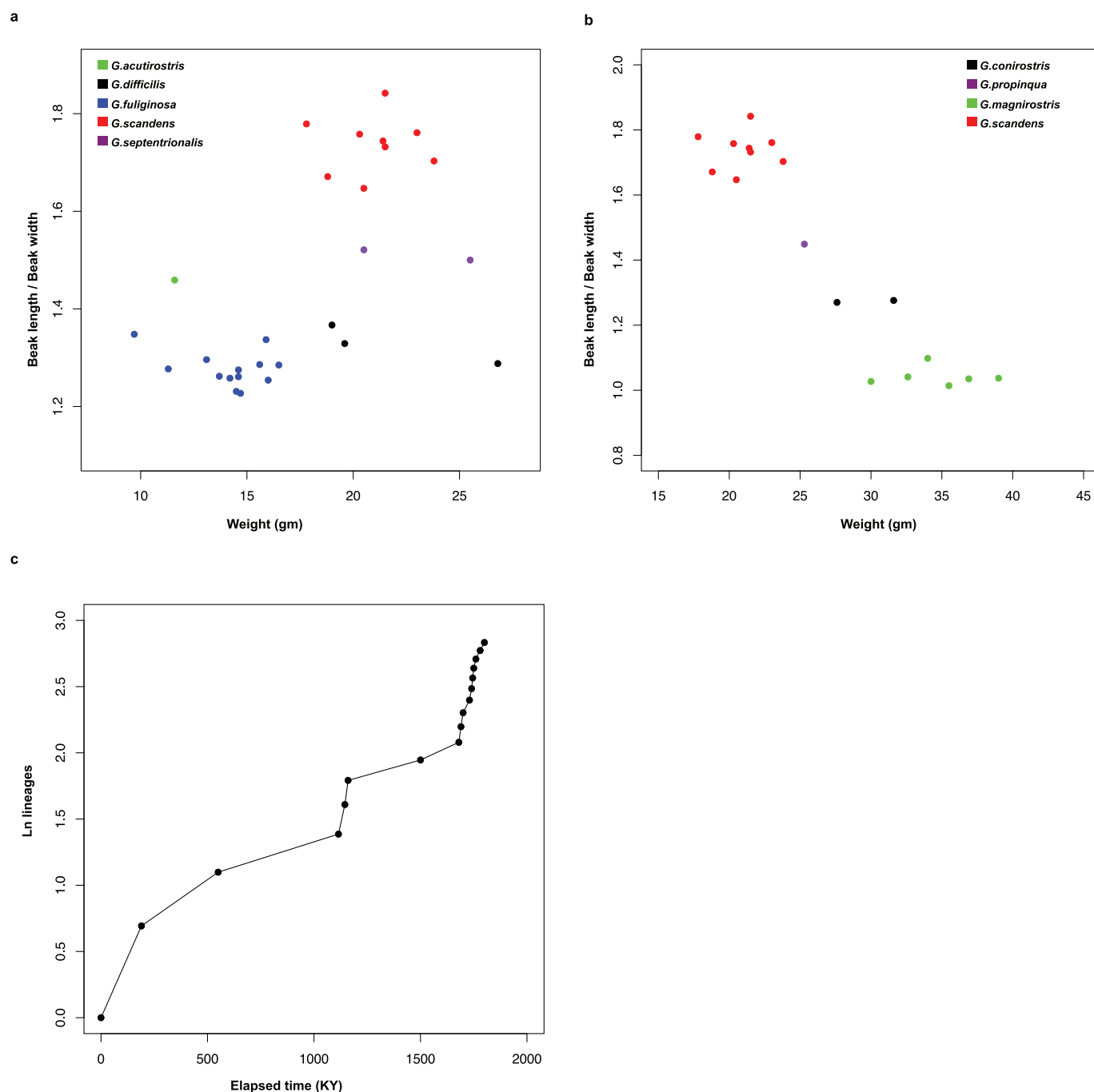


Extended Data Figure 2 | Genetic diversity among Darwin's finches. Heat map illustrating the proportion of shared and fixed polymorphisms among Darwin's finches and outgroup species.



Extended Data Figure 3 | Network tree for the Darwin's finches on the basis of all autosomal sites. Taxa that showed deviations from classical taxonomy are underlined. Finch heads are reproduced from ref. 5. *How and Why*

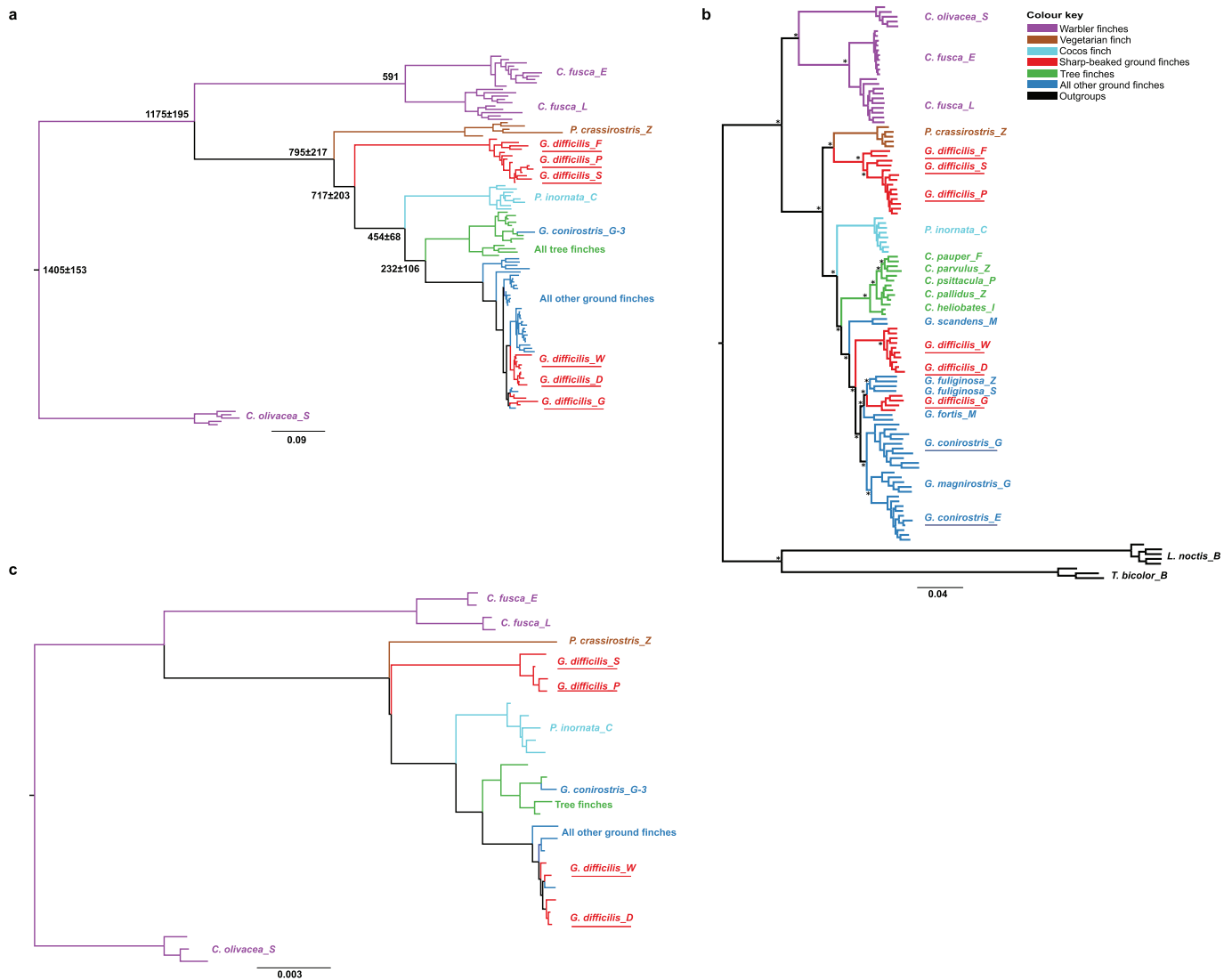
Species Multiply: The Radiation of Darwin's Finches by Peter R. Grant & B. Rosemary Grant. Copyright © 2008 Princeton University Press. Reprinted by permission.



Extended Data Figure 4 | Taxonomy and rate of speciation.

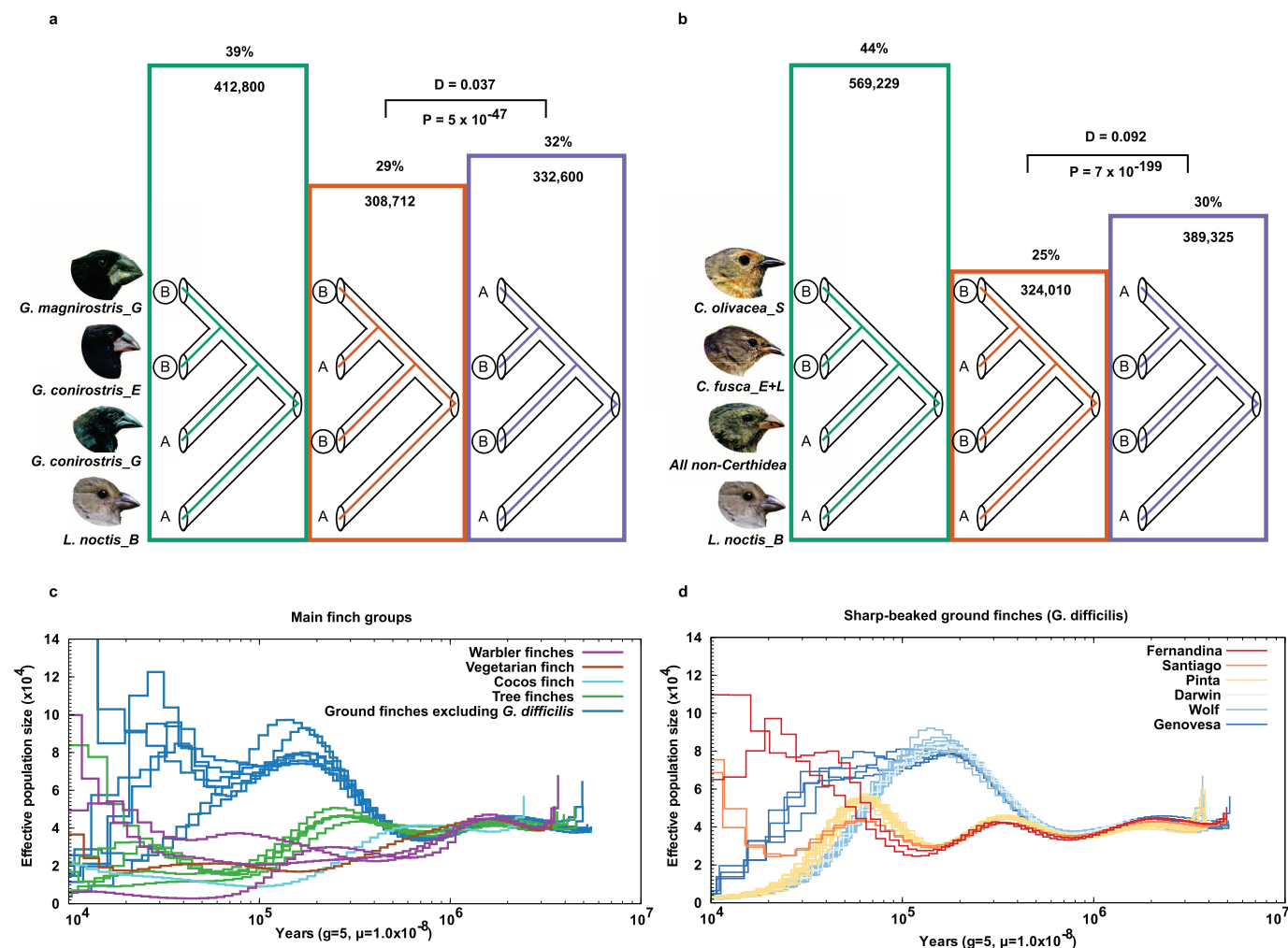
a, Morphological variation among populations of ground finch (*Geospiza*) species, *scandens*, *fuliginosa* and three others, *acutirostris*, *difficilis* and *septentrionalis*, that were formerly classified as a single species (*difficilis*). Data are from refs 56, 57, and from ref. 58 for weights and measures of *difficilis* on Fernandina. **b**, Morphological variation among populations of *G. scandens*, *conirostris*, *propinqua* and *magnirostris* assessed by multiple discriminant function analysis in JMP version 9. In a discriminant function analysis of the measured variables, all populations were correctly identified to species ($-2 \log$

likelihood $P = 0.02$). Maximum discrimination was achieved by entering three variables in the sequence beak width, beak length and body size (weight or wing). Substituting beak depth for beak width gave the same result. No other variable entered significantly. Data are from ref. 57, except for *scandens* and *magnirostris* data from ref. 30. **c**, Species accumulation on a log scale as a function of time before the present, dating based on mtDNA. Species are expected to accumulate linearly according to a 'birth-death' process, eventually declining under a density- (diversity-) dependent mechanism⁵⁹.



Extended Data Figure 5 | Phylogenies for mtDNA and the sex chromosomes Z and W. **a**, Tree based on mtDNA sequences. The dating of the nodes and their variances (in thousands of years) is based on the cytochrome b sequences using the fossil-calibrated divergence rate 2.1% per million years for birds⁴⁶. This tree based on the full mtDNA sequences shows only minor differences compared with previously published trees based only on the

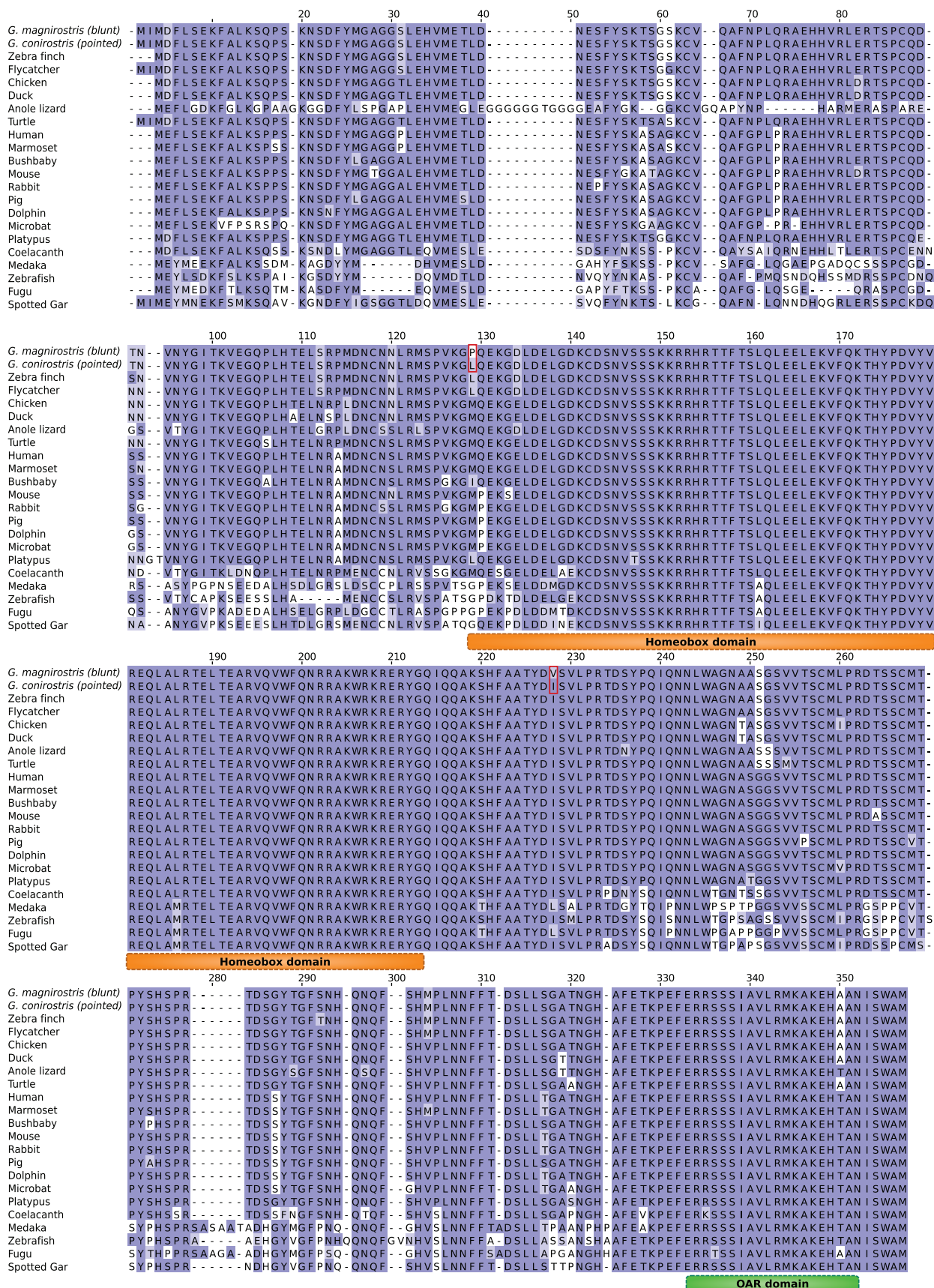
cytochrome b sequence^{6,9}. **b**, Maximum-likelihood trees based on all Z-linked sites; all nodes having full local support on the basis of the Shimodaira–Hasegawa test are marked by asterisks. **c**, Tree based on W sequences, only females. Taxa that showed deviations from classical taxonomy are underscored (applies to **a–c**).



Extended Data Figure 6 | ABBA-BABA analysis and demographic history.

a, ABBA-BABA analysis of *G. magnirostris*, *G. conirostris* on Española and on Genovesa, and with *L. noctis* as outgroup. **b**, Comparison of *C. olivacea*, *C. fusca*, a pool of all non-warblers, and with *L. noctis* as outgroup. The number of informative sites supporting the different trees is indicated both as a

percentage and as the actual number. The D statistic and corresponding Holm-Bonferroni-corrected P value are also given for testing the null hypothesis of symmetry in genetic relationships. Finch heads are reproduced from ref. 5. **c**, PSMC analysis²¹ of all species except the *G. difficilis* group. **d**, PSMC analysis of the *G. difficilis* group.



Extended Data Figure 7 | Sequence conservation of ALX1. Amino-acid alignment of the complete ALX1 sequence among different vertebrates. Amino-acid substitutions between *ALX1* alleles associated with blunt and pointed beaks are highlighted. The homeobox domain is indicated.

Extended Data Table 1 | Phenotypic description of Darwin's finches

Common name	Species	Island	Sampling date (year/month)	Average weight (gm)	Male plumage	Female plumage	Diet*	Beak shape
Large ground finch	<i>Geospiza magnirostris</i>	Genovesa	1989/2	33	Black	Brown, streaked	Seeds	Blunt
Medium ground finch	<i>Geospiza fortis</i>	Daphne Major	1995/1	17	Black	Brown, streaked	Seeds	Blunt
Small ground finch	<i>Geospiza fuliginosa</i>	Santa Cruz	1989/1	13	Black	Brown, streaked	Seeds	Blunt
		Santiago	1996/1	13	Black	Brown, streaked	Seeds	Blunt
Large cactus finch	<i>Geospiza conirostris</i>	Genovesa	1989/2	25	Black	Brown, streaked	Seeds	Pointed
		Española	1997/4	32	Black	Brown, streaked	Seeds	Blunt
Common cactus finch	<i>Geospiza scandens</i>	Daphne Major	2001/4	22	Black	Brown, streaked	Seeds	Pointed
Sharp-beaked ground finch	<i>Geospiza difficilis</i>	Pinta	1997/5	19	Black	Brown, streaked	Insects	Pointed
		Fernandina	1997/4	20	Black	Brown, streaked	Insects	Pointed
		Santiago	1996/1	27	Black	Brown, streaked	Insects	Pointed
		Wolf	1995/1	21	Black	Brown, streaked	Seeds	Pointed
		Darwin	1995/1	25	Black	Brown, streaked	Seeds	Pointed
		Genovesa	1997/4	12	Black	Brown, streaked	Seeds	Pointed
Mangrove finch	<i>Camarhynchus heliobates</i>	Isabela	1998/3	18	Brown-green	Brown-green	Insects	Pointed
Woodpecker finch	<i>Camarhynchus pallidus</i>	Santa Cruz	1998/3	20	Green	Green	Insects	Pointed
Large tree finch	<i>Camarhynchus psittacula</i>	Pinta	1997/5	19	Black and green	Green	Insects	Blunt
Small tree finch	<i>Camarhynchus parvulus</i>	Santa Cruz	1999/2	13	Black and green	Green	Insects	Blunt
Medium tree finch	<i>Camarhynchus pauper</i>	Floreana	1997/4	16	Black and green	Green	Insects	Blunt
Vegetarian finch	<i>Platyspiza crassirostris</i>	Santa Cruz	1988/12	35	Black and brown	Brown, streaked	Fruits	Blunt
Green warbler finch	<i>Certhidea olivacea</i>	Santiago	1996/1	9	Green	Green	Insects	Thin, pointed
Grey warbler finch	<i>Certhidea fusca</i>	Española	1997/4	8	Gray-green	Gray-green	Insects	Thin, pointed
		San Cristóbal	1999/9	8	Gray-green	Gray-green	Insects	Thin, pointed
Cocos finch	<i>Pinaroloxias inornata</i>	Cocos Island	1997/10	16	Black	Brown, streaked	Insects	Thin, pointed

* Primary food type in the dry season when the food is potentially limiting

Extended Data Table 2 | Summary of samples of Darwin's finches and outgroup species

Common name	Species	No. of samples	Island (abbreviation)	Total SNPs*	θ^{**} ($\times 10^{-3}$)	N_e^{***}
Large ground finch	<i>Geospiza magnirostris</i>	5	Genovesa (G)	4,911,160	1.7	41,437
Medium ground finch	<i>Geospiza fortis</i>	2	Daphne Major (M)	3,733,616	2.0	49,222
Small ground finch	<i>Geospiza fuliginosa</i>	2	Santa Cruz (Z)	4,109,669	2.2	54,157
		2	Santiago (S)	4,153,538	2.2	54,563
Large cactus finch	<i>Geospiza conirostris</i>	10	Genovesa (G)	6,530,869	1.8	43,781
		10	Española (E)	5,399,492	1.5	36,221
Common cactus finch	<i>Geospiza scandens</i>	2	Daphne Major (M)	3,272,568	1.8	43,142
Sharp-beaked ground finch	<i>Geospiza difficilis</i>	10	Pinta (P)	3,592,993	1.0	24,109
		2	Fernandina (F)	2,986,435	1.6	39,335
		2	Santiago (S)	2,921,867	1.6	38,039
		8	Wolf (W)	3,184,525	0.9	22,845
		2	Darwin (D)	2,111,758	1.1	27,489
		4	Genovesa (G)	4,652,295	1.8	43,212
Mangrove finch	<i>Camarhynchus heliobates</i>	2	Isabela (I)	1,905,289	1.0	25,115
Woodpecker finch	<i>Camarhynchus pallidus</i>	5	Santa Cruz (Z)	2,805,685	1.0	23,662
Large tree finch	<i>Camarhynchus psittacula</i>	2	Pinta (P)	2,009,269	1.1	26,184
Small tree finch	<i>Camarhynchus parvulus</i>	2	Santa Cruz (Z)	2,595,166	1.4	33,861
Medium tree finch	<i>Camarhynchus pauper</i>	2	Floreana (F)	2,492,881	1.3	32,863
Vegetarian finch	<i>Platyspiza crassirostris</i>	5	Santa Cruz (Z)	2,664,104	0.9	22,491
Green warbler finch	<i>Certhidea olivacea</i>	5	Santiago (S)	2,966,679	1.0	25,047
Grey warbler finch	<i>Certhidea fusca</i>	10	Española (E)	988,062	0.3	6,642
		10	San Cristóbal (L)	4,605,839	1.3	30,931
Cocos finch	<i>Pinaroloxias inornata</i>	8	Cocos Island (C)	2,258,080	0.7	16,232
Black-faced grassquit	<i>Tiaris bicolor</i>	3	Barbados (B)	6,492,110	2.8	69,564
Lesser Antillean bullfinch	<i>Loxigilla noctis</i>	5	Barbados (B)	4,015,128	1.4	34,154

*Total number of polymorphic SNPs within population

** θ = Watterson's theta

*** N_e = Estimated long-term effective population size based on the levels of nucleotide diversity in populations and an estimate of mutation rate of 2.04×10^{-9} per base per year from a comparison between a Darwin's finch and zebra finch (see Supplementary Text).

Architecture of the RNA polymerase II–Mediator core initiation complex

C. Plaschka¹, L. Larivière², L. Wenzek², M. Seizl², M. Hemann², D. Tegunov³, E. V. Petrotchenko⁴, C. H. Borchers⁴, W. Baumeister³, F. Herzog², E. Villa^{3,5} & P. Cramer¹

The conserved co-activator complex Mediator enables regulated transcription initiation by RNA polymerase (Pol) II. Here we reconstitute an active 15-subunit core Mediator (cMed) comprising all essential Mediator subunits from *Saccharomyces cerevisiae*. The cryo-electron microscopic structure of cMed bound to a core initiation complex was determined at 9.7 Å resolution. cMed binds Pol II around the Rpb4–Rpb7 stalk near the carboxy-terminal domain (CTD). The Mediator head module binds the Pol II dock and the TFIIB ribbon and stabilizes the initiation complex. The Mediator middle module extends to the Pol II foot with a ‘plank’ that may influence polymerase conformation. The Mediator subunit Med14 forms a ‘beam’ between the head and middle modules and connects to the tail module that is predicted to bind transcription activators located on upstream DNA. The Mediator ‘arm’ and ‘hook’ domains contribute to a ‘cradle’ that may position the CTD and TFIH kinase to stimulate Pol II phosphorylation.

Transcription initiation at eukaryotic protein-coding genes requires RNA polymerase (Pol) II and the general transcription factors TFIIB, -D, -E, -F, and -H. In the canonical view of initiation^{1,2}, promoter DNA first assembles with TFIIB, the TFIID subunit TBP, and the Pol II–TFIIF complex. The resulting core initiation complex binds TFIIE and TFIIF, which contains an ATPase that unwinds DNA and a kinase that phosphorylates the Pol II carboxy-terminal domain (CTD). RNA synthesis leads to the initially transcribing complex (ITC) and later to release of the general factors and formation of the elongation complex. Recent studies elucidated the three-dimensional architecture of initiation intermediates, including the ITC^{3–9}.

Transcription initiation also requires the co-activator complex Mediator¹⁰. Mediator is recruited by transcription activators, stabilizes the initiation complex, and stimulates TFIIF kinase activity¹¹. Mediator from the yeast *Saccharomyces cerevisiae* contains 25 subunits arranged in four modules. The ‘head’ and ‘middle’ modules are essential for viability, whereas the ‘tail’ and ‘kinase’ modules are not¹². Crystal structures are available for the head module^{13–15} and for subunits in the middle and kinase modules¹². The arrangement of the modules within Mediator was recently revised based on electron microscopy (EM) analysis^{16,17}. EM also visualized endogenous Mediator in complex with Pol II at low resolution^{18–22}. These studies led to inconsistent locations of Mediator on Pol II (Extended Data Fig. 1a), probably owing to Mediator heterogeneity and a lack of general factors.

Here we prepared an active, homogeneous, recombinant 15-subunit core Mediator (cMed) that contains the head and middle modules. cMed corresponds to the endogenous core Mediator purified from yeast²³, and has a human counterpart that was reported while our manuscript was in revision²⁴. cMed binds the core ITC (cITC), which lacks TFIIE and TFIIF and is stabilized by a short six-nucleotide RNA^{6,9}. We resolved the architecture of the cITC–cMed complex by cryo-EM and protein crosslinking. Our results reveal the location of Mediator on Pol II and suggest mechanisms of transcription regulation.

Core Mediator and electron microscopy

We previously prepared Mediator head and middle modules by co-expression of their subunits in bacteria^{14,25}. The head module comprised subunits Med6, Med8, Med11, Med17, Med18, Med20, and Med22. The middle module contained Med4, Med7, Med9, Med10, Med21, and Med31. Recombinant head and middle modules did not bind each other, but their co-expression with subunits Med14 (residues 1–745) and Med19 enabled purification of a stable cMed complex (Methods, Extended Data Fig. 1b). Med14 enabled association of the two modules, consistent with a scaffold function²⁶. cMed comprised 15 conserved²⁷ subunits (Fig. 1a), including all subunits essential for yeast viability¹³. cMed was functional in basal and activated promoter-dependent transcription, whereas the head and middle modules were not (Fig. 1b, Extended Data Fig. 1c).

Pure cMed bound to Pol II and the CTD (Extended Data Fig. 1d–f), and to the cITC (Fig. 1c, d). The cITC–cMed complex was stable during size exclusion chromatography and contained all 31 polypeptides (Fig. 1d, Extended Data Table 1). The sample comprised particles of the expected size when observed by EM in negative stain, and was used for cryo-EM data collection with a direct detection device (Extended Data Fig. 1g, h; Methods). An unbiased reference structure was generated by tomographic reconstruction (Methods). Unsupervised 3D classification of particle images revealed that the cITC–cMed complex co-existed with free cITC and Pol II–DNA/RNA complex (Methods). We obtained high-resolution reconstructions for cITC–cMed, cITC, and Pol II–DNA/RNA complexes at resolutions of 9.7 Å, 7.8 Å, and 6.6 Å, respectively (Figs 2a, 3, Methods).

cITC structure and DNA stabilization

The cITC reconstruction enabled automated fitting of crystal structures of Pol II, TFIIB, TBP, and a TFIIF homology model (Fig. 2a, b; Extended Data Fig. 2a–c; Methods). The remaining EM density belonged to DNA and parts lacking from the TFIIF model (Fig. 2c, d; Extended

¹Max Planck Institute for Biophysical Chemistry, Department of Molecular Biology, Am Fassberg 11, 37077 Göttingen, Germany. ²Gene Center and Department of Biochemistry, Ludwig-Maximilians-Universität München, Feodor-Lynen-Strasse 25, 81377 Munich, Germany. ³Max Planck Institute for Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany. ⁴Department of Biochemistry and Microbiology, Genome British Columbia Protein Centre, University of Victoria, 3101-4464 Markham Street, Victoria, British Columbia V8Z7X8, Canada. ⁵Department of Chemistry and Biochemistry, University of California San Diego, 9500 Gilman Drive, La Jolla, California 92093, USA.

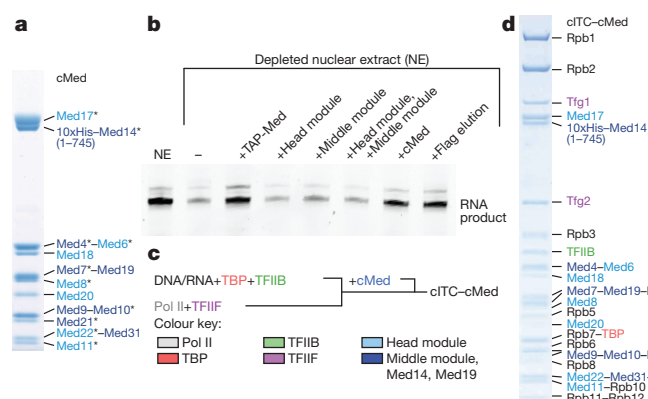


Figure 1 | Reconstitution of Pol II-Mediator complex cITC-cMed. **a**, SDS-PAGE analysis of recombinant 15-subunit cMed. Head and middle module subunits are labelled in blue and violet, respectively. Asterisks mark essential subunits. **b**, Recombinant cMed is functional in promoter-dependent transcription, whereas head and middle modules are not (Methods). As a positive control, Mediator-depleted nuclear extract was complemented with a Mediator fraction, either TAP-Mediator or Flag elution. The RNA product was visualized by primer extension in triplicate experiments. **c**, Assembly of the cITC-cMed complex, and protein colour key used throughout. **d**, SDS-PAGE analysis of the cITC-cMed complex after size exclusion chromatography.

Data Fig. 2d–g). The resulting cITC structure was consistent with a model derived biochemically²⁸, and confirmed and extended our previous model^{3,6,9}. The TFIIB-TBP-DNA complex resided on the Pol II wall, but was slightly rotated towards the Pol II protrusion compared to an earlier model⁹ (Extended Data Fig. 2h, i). The previously unobserved³ TFIIB C-terminal cyclin domain bound upstream DNA and the Pol II protrusion and subunit Rpb12 (Fig. 2).

The cITC reconstruction revealed the path of the DNA template through the Pol II cleft (Fig. 2c, Extended Data Fig. 2g). The template strand descended from the wall to the active site through the

template tunnel¹³ lined by TFIIB and the Pol II fork loop 1 and rudder. At the location expected for the non-template strand²⁹ we observed tubular densities (Fig. 2c). The TFIIF dimerization module resided on the Pol II lobe as in the Pol II-TFIIF complex^{28,30}. The ‘charged helix’ of TFIIF subunit Tfg1 extended from the dimerization module to DNA at position +12 downstream of the transcription start site (Fig. 2d) and may stabilize loaded DNA⁷. The Tfg1 ‘arm’ extended along the Pol II protrusion towards the transcription bubble (Fig. 2d). The linker region of TFIIF subunit Tfg2 passed between the protrusion and TFIIB, and connected to the C-terminal winged helix (WH) domain located at the upstream bubble (Extended Data Fig. 2d, e). The cITC structure is highly conserved⁷, and indicated that the Tfg2 linker and C-terminal winged helix domain stabilize TFIIB and open DNA³¹.

cITC-cMed structure

The cITC model was placed unambiguously into the EM reconstruction of the cITC-cMed complex (Fig. 3, Extended Data Fig. 3a, b). The remaining density belonged to cMed and was 240 Å long, 120 Å high, and 110 Å wide. We observed a single location of cMed on Pol II that differed from previously reported locations (Fig. 3, Extended Data Fig. 1a, Methods). About half of the cMed density corresponded to the head module, and revealed the module’s neck, fixed jaw, and movable jaw^{14,15}. We fitted structures for the two jaws^{14,32} and an improved model for the *S. cerevisiae* neck¹⁵ based on the *Schizosaccharomyces pombe* head module structure¹⁴ (Methods, Extended Data Fig. 3c). The relative position of the neck and jaws was similar in the free module from *S. pombe*¹⁴ but differed in the *S. cerevisiae* structure^{13,14}, indicating mobility (Extended Data Fig. 3d).

The remaining cMed density belonged to the middle module, Med14, and Med19. The density agreed with an extended shape of the middle module²⁵ and was divided into four regions that we call hook, knob, beam, and plank domains (Fig. 3). The middle module was connected to the shoulder and arm of the head module as proposed¹⁴. Additional contacts to the joint and fixed jaw domains in the head module revealed an intimate head-middle module interaction, in agreement with recent data^{16,17}.

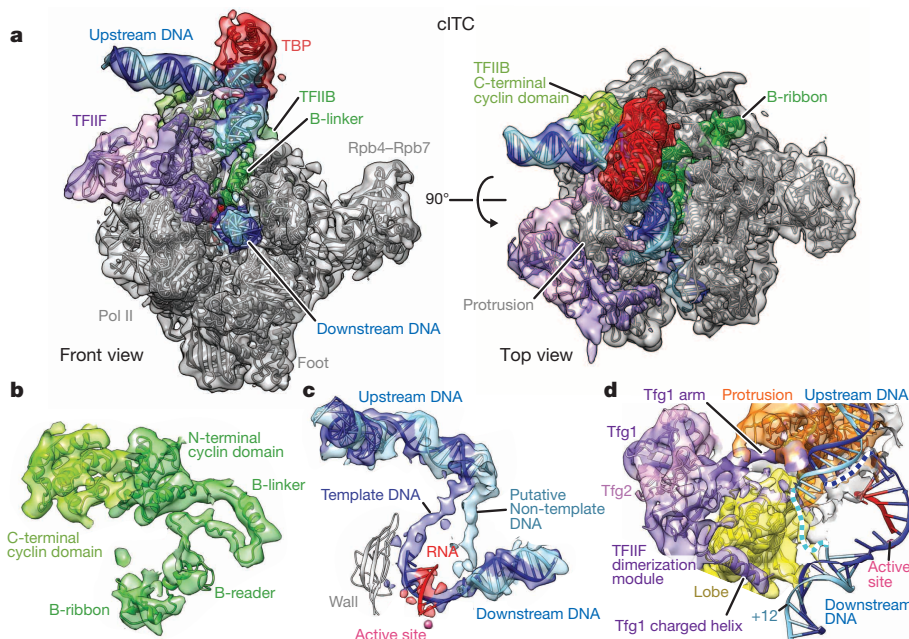


Figure 2 | EM structure of Pol II initiation complex cITC. **a**, EM reconstruction (4,439 particle images, 0.143 FSC = 7.8 Å resolution) with fitted structures as ribbon models. The views are previously defined front and top views of Pol II⁴⁷. The density was filtered by local resolution. **b**, TFIIB X-ray structure³ (ribbon) explains the EM density. View is from the side⁴⁷. **c**, Course of

the DNA template strand through the active centre and putative non-template strand density in the transcription bubble. View is as in **b**. **d**, TFIIF Tfg1 elements. The ‘charged helix’ protrudes towards downstream DNA, whereas the ‘arm’ tracks along the protrusion towards the upstream DNA bubble.

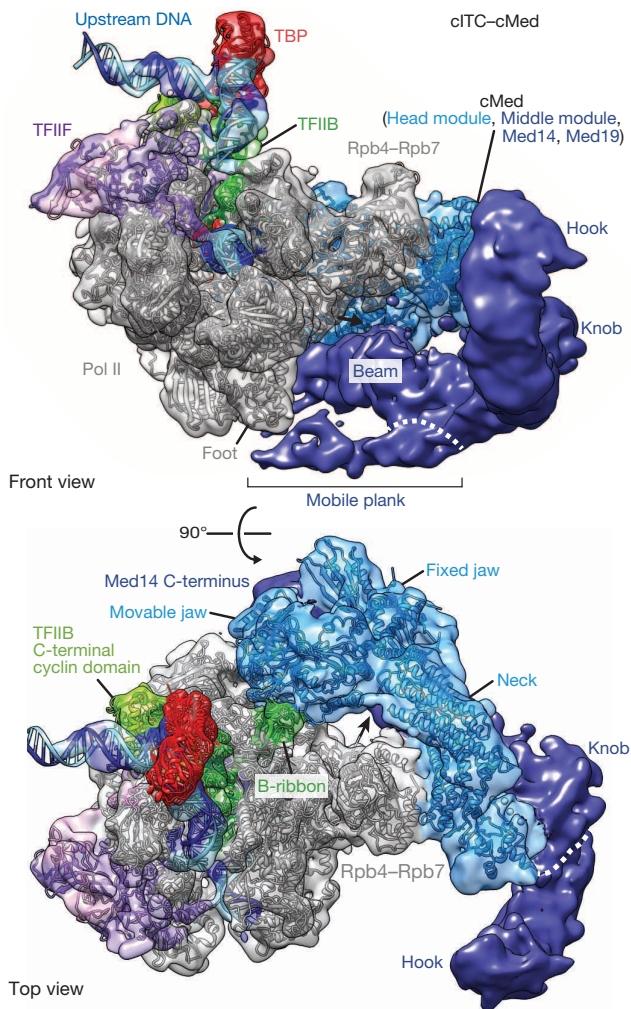


Figure 3 | EM structure of the cITC-cMed complex. EM reconstruction (3,267 particle images, 0.143 FSC = 9.7 Å resolution) viewed as in Fig. 2a with fitted structures (ribbon models). The density was filtered by local resolution. Mediator head and middle modules are coloured in blue and violet, respectively. We observe a single location of cMed on Pol II (compare Extended Data Fig. 8). Head module regions neck, fixed jaw and movable jaw are indicated. A black arrow marks the beginning of the linker to the Pol II CTD.

Mediator topology

Since the middle module is structurally unresolved, we elucidated its subunit topology with protein crosslinking. For the cITC-cMed complex we obtained 706 high-confidence lysine-lysine crosslinks that provided 243 distance restraints (Fig. 4a, Extended Data Fig. 4a–d, Supplementary Table 1). Our structure explained 196 crosslinks within the cITC, 26 crosslinks within the head module, 26 crosslinks between the head module and the remainder of cMed, and 21 crosslinks between cITC and cMed. Crosslinking analysis of free cMed provided 52 additional crosslinks (Supplementary Table 2).

The data indicated that the middle module hook and knob contained subunits Med7, Med10, Med19, Med21, and Med31, and that the globular Med31 domain³³ formed part of the knob (Fig. 4b, Extended Data Fig. 4e). The Med7–Med21 heterodimer resided near the arm and shoulder of the head module. The beam domain contained Med14. The Med14 N-terminal region (residues 1–299) resided between the middle module and the joint and spine domains of the head module. The Med14 C-terminal region (residues 300–745) crosslinked to the head module jaws. The plank contained the Med4–Med9 heterodimer. This resulted in a revised cMed topology that is not consistent with the

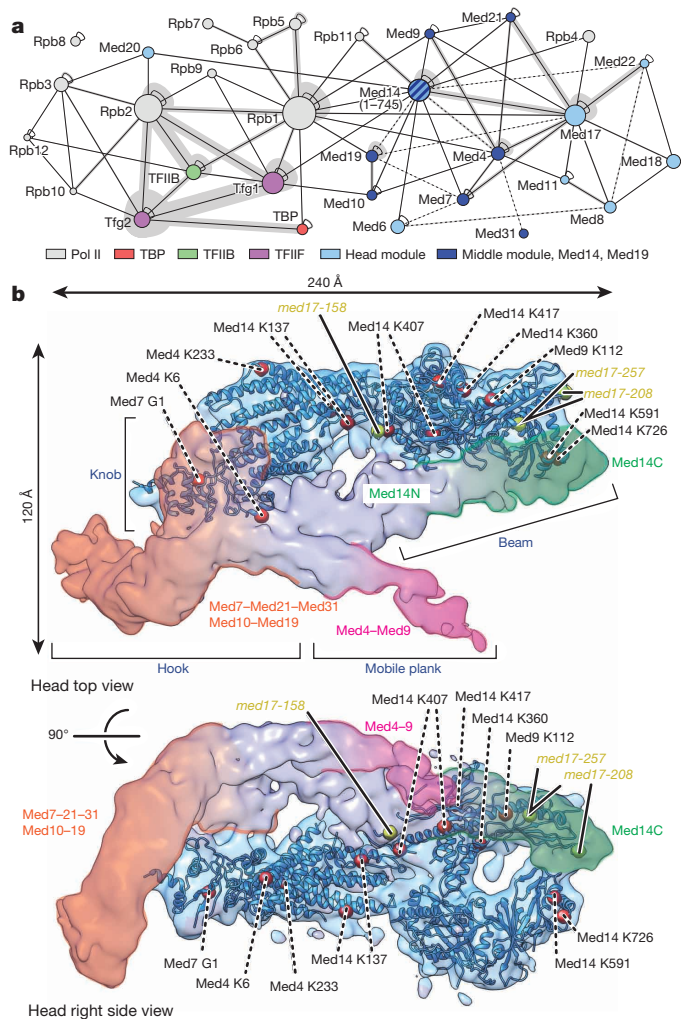


Figure 4 | Protein crosslinking and cMed architecture. **a**, Overview of protein-protein crosslinks visualized with xiNET. Solid and dashed lines indicate crosslinks for cITC-cMed and free cMed, respectively. Circle area correlates with protein mass. Line thickness correlates to the number of crosslinks. 13 crosslinks in flexible regions that exceeded the 27 ± 3 Å distance restraint between C α atoms were excluded. **b**, Architecture of cMed and location of middle module subunits. Views are as defined for the head module¹⁴. Crosslinking residues were mapped onto the head module (red spheres) and labelled with crosslinking residues in the middle module. Mutants med17-158, med17-208, and med17-257 (yellow spheres) that cause synthetic lethality with a Pol II mutation³⁸ map to the head-middle module interface. Putative locations of the Med14 N-terminal (residues 1–299) and C-terminal (residues 300–745) regions are indicated.

canonical Mediator architecture, but with the recent alternative Mediator topology^{16,17} and the architecture of the human cMed counterpart²⁴.

Initiation complex stabilization

In the cITC-cMed structure, cMed binds around the Pol II Rpb4–Rpb7 stalk that is required for initiation³⁴ and forms three interfaces with the cITC (interfaces A to C, Fig. 5). In interface A, the movable jaw heterodimer Med18–Med20 binds the TFIIB B-ribbon domain, the Pol II dock, and the Rpb3–Rpb11 heterodimer (Fig. 5a, b). This contact can explain Mediator-facilitated recruitment of TFIIB³⁵ and a requirement of the B-ribbon for transcription³⁶, and is consistent with SRB mutations in the joint and movable jaw of the head module that rescue the phenotype of the *rpb1-14* mutant that alters the dock³⁷. The location of Med20 between the Rpb3 zinc loop and Rpb3–Rpb11 dimerization domain explains why mutation of the Rpb3 zinc loop impairs Mediator binding³⁸ and is consistent with two Mediator–Pol II crosslinks (Fig. 5b, Extended Data Figs 4b, 5a). The mutant *SRB2-1* comprises a point

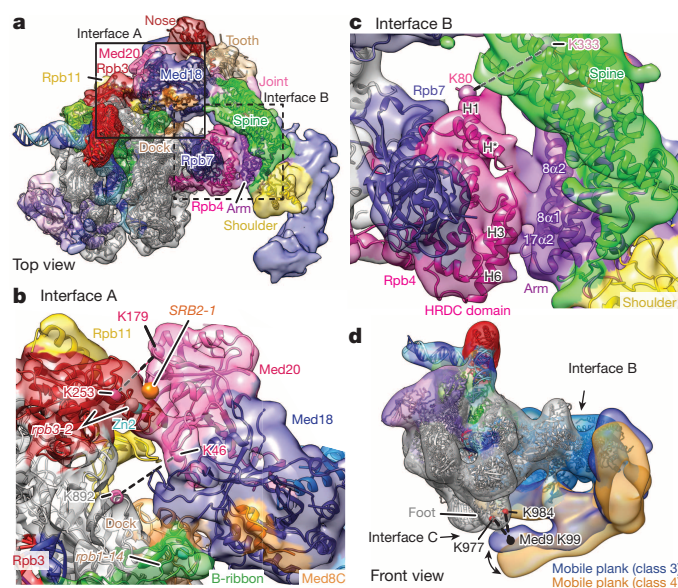


Figure 5 | cITC–cMed interfaces. **a**, View of cITC–cMed complex from the top. Previously defined domains of Pol II⁴⁷ and Mediator head module¹⁴ are in different colours. **b**, Interface A. Spheres depict C α atoms of residues mutated in yeast strains *SRB2-1*, *rpb3-2*, and *rpb1-14*. Crosslinked residues in Pol II and Med20 are indicated. **c**, Interface B. Secondary structure elements are labelled according to the previous nomenclature^{14,40}. A crosslink from Rpb4 to the head module joint is indicated. **d**, Transient interface C. Classification of cITC–cMed particles reveals movement of the plank domain in the middle module. Unsharpened maps at 22 Å resolution for classes 3 and 4 are in violet and gold, respectively (Extended Data Fig. 8b). Crosslinks from the Pol II foot to Med9 K99 (cMed) are indicated. View as in Fig. 3 (top).

mutation in Med20 located in interface A (Fig. 5b), suggesting that weakening this interface suppresses a lethal CTD truncation³⁹. The C-terminal region of Med14 may stabilize interface A because it contacts Med20 and crosslinks to Rpb11 (Extended Data Figs 4c, d, 5a).

Interface B is formed between the conserved arm domain in the head module and the Rpb4–Rpb7 stalk. The interaction involves the arm domain helices $\alpha 1$ and $\alpha 2$ in Med8 and $\alpha 2$ in Med17¹⁴ and α -helices H*, H3, and H6 in Rpb4⁴⁰ (Fig. 5c). The EM density suggests ordering of a mobile insertion between Rpb4 helices H1 and H* and explains a crosslink between H1 and the joint of the head module (Fig. 5c). Interface C is formed between the plank domain in the middle module and the Pol II foot and explains a crosslink from Med9 in the plank to the foot (Fig. 5d). Interface C is transient because the plank is detached from the foot in a subgroup of EM particles (Fig. 5d). The interfaces are hydrophilic and contain conserved residues, suggesting a conserved Pol II–Mediator interaction.

Functional data are consistent with the observed cITC–cMed interfaces. *In vitro*, promoter recruitment of Pol II and TFIIB is lost in extracts from a *srb4-ts* strain, in which the head module dissociates from Mediator^{41,42}. Addition of recombinant head module rescued this defect (Extended Data Fig. 5b), consistent with previous results⁴³ and with our observation that most cITC contacts are formed by the head module. Variants of cMed with amino-terminal truncations of Med8 perturb interface B and were strongly impaired in Pol II binding (Extended Data Fig. 5c). In contrast, cMed variants perturbing either interface A (lacking the movable jaw) or interface C (lacking Med4–Med9) still bound Pol II, showing that interface B is critical for cMed binding to Pol II. Consistent with these findings, perturbation of interface A *in vivo* by nuclear depletion of Med18 led to a mild decrease in messenger RNA synthesis rates⁴⁴, whereas perturbation of interface B by deletion of Rpb4 or dissociation of the head module (interfaces A and B) resulted in a strong, global reduction of mRNA synthesis (Extended Data Fig. 6 and ref. 45).

Pol II activation and phosphorylation

Transcription activators located on upstream DNA can recruit Mediator by binding its tail module. Consistent with this, the tail module is predicted to reside near upstream DNA as shown by superposition of the EM density of free Mediator¹⁷ onto our structure (Extended Data Fig. 7a). The superposition further argues that Mediator does not undergo major structural changes upon Pol II binding. Transcription activation may however include conformational control of Pol II. We propose that cMed stabilizes an active arrangement⁴⁶ of the two major polymerase modules⁴⁷ ‘core’ and ‘shelf’, because it bridges between these modules via interfaces A and C, respectively. Mediator may also communicate with TFIIS allosterically (Extended Data Fig. 7b), because TFIIS reorients the shelf module⁴⁸ that contains the foot, and cooperates with Mediator during initial transcription⁴⁹. Mediator may also facilitate DNA opening by positioning the clamp and the Rpb4–Rpb7 stalk (Extended Data Fig. 7c).

We considered how Mediator may stimulate CTD phosphorylation at serine-5 residues by TFIH¹¹. Comparison of our structure with the EM reconstruction of the human initiation complex that includes TFIIE and the core of TFIH⁷ suggested that TFIIE reaches the head module arm and shoulder, and that the TFIH core may bind the middle module (Fig. 6a, Extended Data Fig. 7c, d). This may explain how Mediator stabilizes TFIIE and TFIH in the initiation complex⁵⁰. The TFIH kinase subcomplex, which was modelled onto the human complex⁷, projects into a ‘cradle’ formed between Pol II and the Mediator arm and hook domains (Fig. 6a, Extended Data Fig. 7b, c). A part of the CTD was apparently ordered inside the cradle on the surface of the arm domain at a location similar to that of a CTD peptide observed in head module crystals¹⁵ (Fig. 6b, Extended Data Fig. 7e). Thus Mediator may stimulate CTD phosphorylation by orienting the mobile TFIH kinase and CTD within the cradle, consistent with a previous suggestion¹¹. Indeed, cMed could stimulate CTD phosphorylation by TFIH *in vitro* (Extended data Fig. 7f), although less strongly than endogenous Mediator (not shown).

Conclusions

The structure of the yeast cITC reveals a conserved architecture of the core transcription initiation complex in eukaryotic cells and elucidates functions of general transcription factors. Preparation of an active, recombinant cMed complex led to the cITC–cMed structure, which supports a revised architecture of Mediator and reveals the location of cMed on Pol II. The results suggest how Mediator binds transcription activators located on upstream DNA, how it stabilizes the

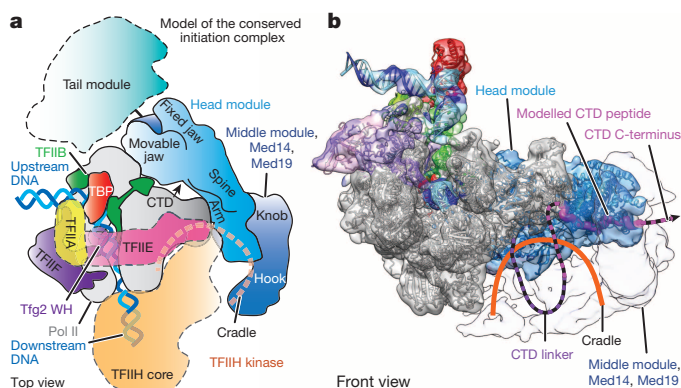


Figure 6 | Initiation complex model and CTD cradle. **a**, Schematic of the Pol II–Mediator initiation complex viewed from the top⁴⁷. Locations of TFIIE, TFIIE and TFIH were inferred from the EM reconstruction of the human initiation complex⁷ (Extended Data Fig. 7d). The location of the Mediator tail module was inferred from the revised Mediator architecture¹⁷ (Extended Data Fig. 7a). **b**, View from the front into the ‘cradle’ of the cITC–cMed complex. A modelled CTD peptide (magenta) was positioned according to the head module–CTD structure¹⁵ (PDB code 4GWQ). The middle module is transparent.

initiation complex, and how it stimulates Pol II activation and CTD phosphorylation.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 13 June 2014; accepted 14 January 2015.

Published online 4 February 2015.

- Buratoski, S., Hahn, S., Guarente, L. & Sharp, P. Five intermediate complexes in transcription initiation by RNA polymerase II. *Cell* **56**, 549–561 (1989).
- Hahn, S. & Young, E. T. Transcriptional regulation in *Saccharomyces cerevisiae*: transcription factor regulation and function, mechanisms of initiation, and roles of activators and coactivators. *Genetics* **189**, 705–736 (2011).
- Kostrewa, D. *et al.* RNA polymerase II-TFIIB structure and mechanism of transcription initiation. *Nature* **462**, 323–330 (2009).
- Liu, X., Bushnell, D., Wang, D., Calero, G. & Kornberg, R. D. Structure of an RNA polymerase II-TFIIB complex and the transcription initiation mechanism. *Science* **327**, 206–209 (2010).
- Grünberg, S., Warfield, L. & Hahn, S. Architecture of the RNA polymerase II preinitiation complex and mechanism of ATP-dependent promoter opening. *Nature Struct. Mol. Biol.* **19**, 788–796 (2012).
- Sainsbury, S., Niesser, J. & Cramer, P. Structure and function of the initially transcribing RNA polymerase II-TFIIB complex. *Nature* **493**, 437–440 (2013).
- He, Y., Fang, J., Taatjes, D. & Nogales, E. Structural visualization of key steps in human transcription initiation. *Nature* **495**, 481–486 (2013).
- Murakami, K. *et al.* Architecture of an RNA polymerase II transcription pre-initiation complex. *Science* **342**, 1238724 (2013).
- Mühlbacher, W. *et al.* Conserved architecture of the core RNA polymerase II initiation complex. *Nature Commun.* **5**, 4310 (2014).
- Conaway, R. C. & Conaway, J. Origins and activity of the Mediator complex. *Semin. Cell Dev. Biol.* **22**, 729–734 (2011).
- Max, T., Søgaard, M. & Svejstrup, J. Q. Hyperphosphorylation of the C-terminal repeat domain of RNA polymerase II facilitates dissociation of its complex with mediator. *J. Biol. Chem.* **282**, 14113–14120 (2007).
- Larivière, L., Seizl, M. & Cramer, P. A structural perspective on Mediator function. *Curr. Opin. Cell Biol.* **24**, 305–313 (2012).
- Imasaki, T. *et al.* Architecture of the Mediator head module. *Nature* **475**, 240–243 (2011).
- Larivière, L. *et al.* Structure of the Mediator head module. *Nature* **492**, 448–451 (2012).
- Robinson, P. J., Bushnell, D., Trnka, M., Burlingame, A. & Kornberg, R. Structure of the mediator head module bound to the carboxy-terminal domain of RNA polymerase II. *Proc. Natl Acad. Sci. USA* **109**, 17931–17935 (2012).
- Wang, X. *et al.* Redefining the modular organization of the core Mediator complex. *Cell Res.* **24**, 796–808 (2014).
- Tsai, K.-L. *et al.* Subunit architecture and functional modular rearrangements of the transcriptional mediator complex. *Cell* **157**, 1430–1444 (2014).
- Asturias, F. J., Jiang, Y. W., Myers, L. C., Gustafson, C. M. & Kornberg, R. D. Conserved structures of Mediator and RNA polymerase II holoenzyme. *Science* **283**, 985–987 (1999).
- Davis, J. A., Takagi, Y., Kornberg, R. & Asturias, F. Structure of the yeast RNA polymerase II holoenzyme: Mediator conformation and polymerase interaction. *Mol. Cell* **10**, 409–415 (2002).
- Elmlund, H. *et al.* The cyclin-dependent kinase 8 module sterically blocks Mediator interactions with RNA polymerase II. *Proc. Natl Acad. Sci. USA* **103**, 15788–15793 (2006).
- Bernecky, C., Grob, P., Ebmeier, C., Nogales, E. & Taatjes, D. Molecular architecture of the human Mediator-RNA polymerase II-TFIIF assembly. *PLoS Biol.* **9**, e1000603 (2011).
- Tsai, K.-L. *et al.* A conserved Mediator-CDK8 kinase module association regulates Mediator-RNA polymerase II interaction. *Nature Struct. Mol. Biol.* **20**, 611–619 (2013).
- Liu, Y., Ranish, J., Aebersold, R. & Hahn, S. Yeast nuclear extract contains two major forms of RNA polymerase II mediator complexes. *J. Biol. Chem.* **276**, 7169–7175 (2001).
- Cevher, M. A. *et al.* Reconstitution of active human core Mediator complex reveals a critical role of the MED14 subunit. *Nature Struct. Mol. Biol.* **21**, 1028–1034 (2014).
- Larivière, L. *et al.* Model of the Mediator middle module based on protein cross-linking. *Nucleic Acids Res.* **41**, 9266–9273 (2013).
- Lee, Y. C., Park, J., Min, S., Han, S. & Kim, Y. An activator binding module of yeast RNA polymerase II holoenzyme. *Mol. Cell Biol.* **19**, 2967–2976 (1999).
- Bourbon, H.-M. Comparative genomics supports a deep evolutionary origin for the large, four-module transcriptional mediator complex. *Nucleic Acids Res.* **36**, 3993–4008 (2008).
- Eichner, J., Chen, H.-T., Warfield, L. & Hahn, S. Position of the general transcription factor TFIIF within the RNA polymerase II transcription preinitiation complex. *EMBO J.* **29**, 706–716 (2010).
- Zhang, Y. *et al.* Structural basis of transcription initiation. *Science* **338**, 1076–1080 (2012).
- Chen, Z. A. *et al.* Architecture of the RNA polymerase II-TFIIF complex revealed by cross-linking and mass spectrometry. *EMBO J.* **29**, 717–726 (2010).
- Fishburn, J. & Hahn, S. Architecture of the yeast RNA polymerase II open complex and regulation of activity by TFIIF. *Mol. Cell Biol.* **32**, 12–25 (2012).
- Larivière, L. *et al.* Structure and TBP binding of the Mediator head subcomplex Med8–Med18–Med20. *Nature Struct. Mol. Biol.* **13**, 895–901 (2006).
- Koschubs, T. *et al.* Identification, structure, and functional requirement of the Mediator submodule Med7N/31. *EMBO J.* **28**, 69–80 (2009).
- Edwards, A. M., Kane, C. M., Young, R. A. & Kornberg, R. D. Two dissociable subunits of yeast RNA polymerase II stimulate the initiation of transcription at a promoter *in vitro*. *J. Biol. Chem.* **266**, 71–75 (1991).
- Baek, H. J., Kang, Y. & Roeder, R. Human Mediator enhances basal transcription by facilitating recruitment of transcription factor IIB during preinitiation complex assembly. *J. Biol. Chem.* **281**, 15172–15181 (2006).
- Ranish, J. A., Yudkovsky, N. & Hahn, S. Intermediates in formation and activity of the RNA polymerase II preinitiation complex: holoenzyme recruitment and a postrecruitment role for the TATA box and TFIIB. *Genes Dev.* **13**, 49–63 (1999).
- Thompson, C. M., Koleske, A., Chao, D. & Young, R. A. Multisubunit complex associated with the RNA polymerase II CTD and TATA-binding protein in yeast. *Cell* **73**, 1361–1375 (1993).
- Soutourina, J., Wydau, S., Ambroise, Y., Boschiero, C. & Werner, M. Direct interaction of RNA polymerase II and mediator required for transcription *in vivo*. *Science* **331**, 1451–1454 (2011).
- Nonet, M. L. & Young, R. Intragenic and extragenic suppressors of mutations in the heptapeptide repeat domain of *Saccharomyces cerevisiae* RNA polymerase II. *Genetics* **123**, 715–724 (1989).
- Armache, K. J., Mitterweger, S., Meinhardt, A. & Cramer, P. Structures of complete RNA polymerase II and its subcomplex, Rpb4/7. *J. Biol. Chem.* **280**, 7131–7134 (2005).
- Linder, T., Zhu, X., Baraznenok, V. & Gustafsson, C. The classical *srb4-138* mutant allele causes dissociation of yeast Mediator. *Biochem. Biophys. Res. Commun.* **349**, 948–953 (2006).
- Takagi, Y. & Kornberg, R. D. Mediator as a general transcription factor. *J. Biol. Chem.* **281**, 80–89 (2006).
- Takagi, Y. *et al.* Head module control of mediator interactions. *Mol. Cell* **23**, 355–364 (2006).
- Sun, M. *et al.* Comparative dynamic transcriptome analysis (cDTA) reveals mutual feedback between mRNA synthesis and degradation. *Genome Res.* **22**, 1350–1359 (2012).
- Schulz, D., Pirkil, N., Lehmann, E. & Cramer, P. Rpb4 subunit functions mainly in mRNA synthesis by RNA polymerase II. *J. Biol. Chem.* **289**, 17446–17452 (2014).
- Engel, C., Sainsbury, S., Cheung, A., Kostrewa, D. & Cramer, P. RNA polymerase I structure and transcription regulation. *Nature* **502**, 650–655 (2013).
- Cramer, P., Bushnell, D. & Kornberg, R. Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution. *Science* **292**, 1863–1876 (2001).
- Kettenberger, H., Armache, K.-J. & Cramer, P. Architecture of the RNA polymerase II-TFIIS complex and implications for mRNA cleavage. *Cell* **114**, 347–357 (2003).
- Nock, A., Ascano, J., Barrero, M. & Malik, S. Mediator-regulated transcription through the +1 nucleosome. *Mol. Cell* **48**, 837–848 (2012).
- Esnault, C. *et al.* Mediator-dependent recruitment of TFIIF modules in preinitiation complex. *Mol. Cell* **31**, 337–346 (2008).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank C. Bernecky, W. Mühlbacher, B. Schwalb, P. Eser, S. Etzold from the Cramer laboratory and J. Plitzko and O. Mihalache from the Baumeister laboratory for help. We thank J. Schuller, S. Pfeffer and F. Förster for help with negative-stain tomography data acquisition and reconstruction of the initial reference. We thank N. Sessler and K. Makepeace from the Borchers laboratory for help with mass spectrometry. We thank M. Raabe and H. Urlaub for protein identification. We thank S. Hahn for providing *Srb5* (3 × Flag) and TFIIF (TAP-Rad3) yeast strains. M.S. was supported by a Boehringer Ingelheim fellowship and the Elite Network of Bavaria. C.H.B. was supported by Genome Canada and Genome British Columbia Science and Technology Innovation Centre funding and The Natural Sciences and Engineering Research Council of Canada grant. F.H. was supported by grants of the LMUexcellent initiative, the Bavarian Research Center of Molecular Biosystems and the Deutsche Forschungsgemeinschaft/GRK1721. P.C. was supported by the Deutsche Forschungsgemeinschaft grants SFB646 and GRK1721 (the latter to P.C. and E.V.), the European Research Council Advanced Grant TRANSIT, the Jung-Stiftung, and the Volkswagen Foundation.

Author Contributions C.P. designed and carried out all experiments, except for the following. L.L. performed initial cloning and purification of cMed and prepared the *S. cerevisiae* head module homology model. L.W. assisted with protein purification and functional assays. M.S. carried out immobilized template and cDTA assays. M.H. carried out mass spectrometry of crosslinked cITC–cMed peptides. F.H. supervised mass spectrometry of cITC–cMed. E.V.P. and C.H.B. supervised mass spectrometry of cMed. W.B. provided essential materials and expertise. E.V. helped with EM data collection and supervised electron microscopy. E.V. and D.T. helped with EM data processing. P.C. designed and supervised research. C.P. and P.C. interpreted the data and wrote the manuscript.

Author Information 3D cryo-EM density maps of Pol II–DNA/RNA, cITC, and cITC–cMed have been deposited in the Electron Microscopy Database under accession numbers EMD-2784, EMD-2785, and EMD-2786, respectively. Coordinate files of Pol II–DNA/RNA, cITC, and cITC–cMed models have been deposited in the Protein Data Bank under accession numbers 4V1M, 4V1N, and 4V1O. Microarray data were deposited in ArrayExpress under accession number E-MTAB-2942. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to P.C. (pcramer@mpibpc.mpg.de).

METHODS

Vectors and sequences. Vectors for co-expression of *Saccharomyces cerevisiae* (Sc) core Mediator (cMed) subunits in *Escherichia coli* are shown in Extended Data Fig. 1b. Open reading frames (ORFs) of Med19 and Med14 (1–745; ref. 51) with an additional N-terminal 10×histidine tag were cloned sequentially into a pET Duet vector (Novagen). The histidine tags of previously cloned genes were removed^{14,25}. Ribosomal binding sites were introduced as described⁵². Sequences are available upon request. All proteins were expressed in *E. coli* BL21(DE3)RIL (Stratagene). Mediator head and middle modules were expressed and purified as described^{14,25}.

Preparation of core Mediator. For co-expression of the 15-subunit cMed, *E. coli* BL21(DE3)RIL cells were transformed with three plasmids (Extended Data Fig. 1b) and grown in LB medium at 37 °C to an optical density of 0.7 at 600 nm. Expression was induced with 0.5 mM isopropyl-β-D-thiogalactoside (IPTG) for 24 h at 18 °C. Cells were harvested and resuspended in buffer A (50 mM HEPES-KOH pH 7.5, 400 mM KCl, 10% glycerol, 10 mM imidazole, 5 mM dithiothreitol (DTT)) containing protease inhibitors⁵³. After sonication and centrifugation, the supernatant was loaded on a HisTrap HP 5 ml column (GE Healthcare) equilibrated in buffer B (25 mM HEPES-KOH pH 7.5, 400 mM KCl, 10% glycerol, 25 mM imidazole, 5 mM DTT). The column was washed with seven column volumes (CVs) of buffer B. The complex was eluted with a linear gradient from 25 mM to 300 mM imidazole in buffer B over 10 CVs. Fractions containing the complex were diluted 1:3 in buffer C (25 mM HEPES-KOH pH 7.5, 100 mM KCl, 10% glycerol, 1 mM EDTA, 5 mM DTT) and applied to a HiTrap Q HP 5 ml column (GE Healthcare) equilibrated in buffer C. The complex was eluted with a gradient from 100 mM to 800 mM KCl in buffer C over 25 CVs. Fractions containing the complex were concentrated and applied to gel filtration using a Superose 6 10/600 GL (GE Healthcare) equilibrated in buffer D (25 mM HEPES-KOH pH 7.5, 400 mM KCl, 3 mM DTT). Purified cMed was concentrated to 3 mg ml⁻¹, flash-frozen and stored at -80 °C. Up to 2 mg of pure cMed could be obtained from 8 l cell culture (Fig. 1a). The identity of the protein subunits was confirmed by mass spectrometry.

Promoter-dependent *in vitro* transcription. Nuclear extract (NE) from 3 l yeast culture of strain SHY349 was prepared as described^{53,54}. SHY349 NE was immunodepleted of endogenous Mediator as described⁵⁴, with minor modifications. 150 μl NE were dialysed for 90 min against buffer E (20 mM HEPES-KOH pH 7.6, 75 mM ammonium sulphate, 10 mM MgSO₄, 20% glycerol, 1 mM EGTA pH 8.0, 0.5 mM DTT) containing protease inhibitors⁵³. 75 μl anti-Flag M2 agarose beads (Sigma) were washed twice with buffer E, and incubated in NE for 1 h at 4 °C on a turning wheel. After centrifugation, the supernatant was depleted with the same amount of beads used initially. Beads were removed by centrifugation. The doubly depleted NE was flash-frozen and stored at -80 °C. Tandem affinity purification (TAP) of endogenous Mediator was performed using a C-terminal TAP tag on Med7 as described⁵³. Purified TAP-Mediator was flash-frozen and stored at -80 °C. Activator-dependent and independent promoter-dependent *in vitro* transcription and primer extension was performed as described⁵³. Purified TAP-Mediator (~0.25 pmol), 7-subunit head module (1.25 pmol), 6-subunit middle module (1.25 pmol) or cMed (2.5 pmol) were added as indicated in Fig. 1b.

Preparation of cITC–cMed complex. *S. cerevisiae* 12-subunit Pol II⁵⁵, TBP (residues 61–240)⁵⁶, TFIIB⁶, and TFIIF (*Saccharomyces mikatae* Tfg1, *S. cerevisiae* Tfg2)⁵⁷ were prepared as described. The nucleic acid scaffold previously used to generate the core Pol II initially transcribing complex (cITC)⁹, was used for assembly of the cITC–cMed complex (template 5′-CGAGAACAGTAGCACGC TGTGTATATAATAGTGTGTGTACATAGCGGAGGTCGGTGGGGCACAA CTGCGCT-3′; nontemplate, 5′-AGCGAGTGTGTCTATGATATTTTATGT ATGTACAACACACTATTATACACGCGTGTCTACTGTCTCTCG-3′; RNA, 5′-AUAUCA-3′). Pol II (150 μg at 3 mg ml⁻¹) was incubated with a fourfold molar excess of TFIIF for 5 min at 25 °C. A twofold molar excess of nucleic acid scaffold and a fourfold molar excess of TBP and TFIIB were added to buffer F (25 mM HEPES-KOH pH 7.5, 180 mM potassium acetate, 5% glycerol, 5 mM DTT) and incubated with pre-formed Pol II–TFIIF complex for 10 min at 25 °C. cMed was added to the cITC in a 1.2-fold molar excess over Pol II and incubated for 50 min at 25 °C. The sample was cooled in 5 min intervals from 25 °C to 20 °C, 15 °C, 10 °C and 4 °C. The cITC–cMed complex was purified by gel filtration using a Superose 6 10/600 GL equilibrated in buffer F. Fractions containing the complex were concentrated to ~0.6 mg ml⁻¹ and extra nucleic acid scaffold was added in equimolar amount.

Binary interaction assays. To test the binary interaction of cMed with glutathione-S-transferase (GST)–CTD, GST and GST–CTD were expressed separately in *E. coli* BL21(DE3)RIL cells and induced at an optical density of 0.5 at 600 nm with 0.5 mM IPTG. Cells were lysed by sonication in buffer G (20 mM HEPES-KOH pH 7.5, 100 mM KCl, 10% glycerol, 100 μM EDTA, 0.1% NP40, 1 mM PMSF, 5 mM DTT). After centrifugation, the supernatant was applied to 1 ml glutathione Sepharose 4B resin (GE Healthcare), equilibrated in buffer G. The resin was washed twice with 10 ml buffer G. 10 μg purified cMed were incubated for 3 h

on ice with 30 μl immobilized GST or GST–CTD resin. The resin was washed four times with 600 μl buffer G and bound proteins were analysed by SDS–PAGE.

To test the binary Pol II–cMed interaction by protein pull-down, 3 μg purified Pol II was biotinylated on the Rpb3 subunit as described⁵⁸ and immobilized on 20 μl Dynabeads M280 Streptavidin resin (Life Technologies), equilibrated in buffer F. 1.5 μg cMed were incubated with immobilized Pol II or control beads for 1 h at 4 °C. Beads were washed four times and bound proteins were analysed by SDS–PAGE. To further assess the stability of Pol II–cMed by gel filtration, 80 μg Pol II were incubated with 1.5-fold less cMed as done for the cITC–cMed complex (see Preparation of cITC–cMed complex). Pol II–cMed complex was applied to size exclusion chromatography on a Superose 6 10/600 GL equilibrated in buffer F, and peak fractions were analysed by SDS–PAGE.

To test the interaction of Pol II with cMed variants comprising mutations in interface A, B and C, 2 μg recombinant cMed variant were incubated with 3 μg biotinylated Pol II immobilized on 15 μl Dynabeads M280 Streptavidin resin (Life Technologies) equilibrated in buffer F. Beads were washed four times and bound proteins were analysed by SDS–PAGE. cMed variants were purified according to the protocol for complete cMed, yet gave much lower yield.

Crosslinking and mass spectrometry. 80 μg of purified cITC–cMed were crosslinked⁵⁹ with 1–4 mM di-succinimidyl-suberate (DSS-d0/d12, Creative Molecules Inc.) or 1–4 mM bis-sulfo-succinimidyl-suberate (BS3-d0/d12, Creative Molecules Inc.) as described⁶⁰. Crosslinked samples were digested with trypsin or AspN. Crosslinked peptides were enriched and analysed on a liquid chromatography system coupled to the electrospray ionization (ESI) source of an Orbitrap Elite mass spectrometer (Thermo Scientific)⁶⁰. Crosslinks were identified by xQuest⁶¹ as described⁶⁰. 35 μg of purified cMed were crosslinked with 0.1–0.15 mM cyanuribiotin-dipropionyl succinimide (CBDPS-d0/d8, Creative Molecules Inc.) as described²⁵. Crosslinked cMed was digested with trypsin and/or GluC, peptides were enriched and analysed on a liquid chromatography system coupled to the ESI source of an LTQ Orbitrap Velos mass spectrometer (Thermo Scientific) as described²⁵. Crosslinks were identified by DXMSMS Match of ICC-CLASS⁶².

Electron microscopy. Purified cITC–cMed complex was crosslinked with 2 mM BS3 (Sigma) for 30 min at 30 °C, and the reaction quenched with 50 mM ammonium bicarbonate. The crosslinked sample was purified by gel filtration using a Superose 6 10/600 GL equilibrated in buffer F. Fractions containing the complex were concentrated to ~0.15 mg ml⁻¹. Negatively stained samples of the cITC–cMed were prepared on continuous carbon coated grids (Quantifoil). Grids were glow-discharged 30 s before deposition of 4 μl sample (~0.15 mg ml⁻¹) and incubated for 1 min. Grids were blotted between sequential transfers to two 40 μl drops distilled water, stained for 1 min in a 40 μl drop 2% (w/v) uranyl acetate solution, and blotted until dry. To obtain an unbiased initial model, a 3D reconstruction was generated from particles selected from negative-stain tomography data. Single axis tilt series were recorded using Serial EM⁶³ on a FEI Tecnai F20 microscope operated at 200 keV. Images were acquired from -54° to 54° with an angular increment of 3°. Images were recorded on a 4k×4k Gatan Ultrascan CCD camera with a defocus of -2 μm and at a nominal magnification of 68,000× (2.21 Å pixel⁻¹). The cumulative dose per tomogram did not exceed 100 e⁻ Å⁻². To refine the negative-stain tomography reconstruction, a further 103 micrographs of the untitled sample were acquired with a range of defocus values (from -0.5 μm to -1.5 μm) and used for single-particle analysis.

Cryo samples of the cITC–cMed were prepared on lacey carbon copper grids (Quantifoil). Grids were glow-discharged for 20 s before deposition of 4 μl sample (~0.15 mg ml⁻¹) and incubated for 30 s. Grids were washed twice with 4 μl distilled water, blotted, and vitrified by plunging into liquid ethane with a manual plunger. Data was acquired using the TOM toolbox⁶⁴ on a FEI Titan Krios operated in EFTEM mode at 300 keV. 2,972 movies were collected using a Gatan K2 Summit direct detector with a range of defocus values (from -1 μm to -2.5 μm) at a nominal magnification of 37,000× (1.35 Å pixel⁻¹). The camera was operated in 'super-resolution' mode (0.675 Å pixel⁻¹) with exposure times of 0.2–0.3 s per frame, a dose rate of ~8 e⁻ pixel⁻¹ s⁻¹, and a target total dose of 25–30 e⁻ Å⁻². Movies were binned once in Fourier space, and partitioned into 2,048² quadrants that were aligned and averaged using a CUDA implementation of a previously described algorithm⁶⁵. The averaged 2,048² images were used for image processing.

Image processing. Negative stain tomography data was processed using the TOM toolbox⁶⁴. The tilt series was contrast-transfer function (CTF) corrected as described⁶⁶. Due to the absence of colloidal markers, images were aligned using feature tracking before weighted back projection. 175 subvolumes were selected manually using EMAN2⁶⁷ and extracted with a 160³ voxel box size from the reconstructed tomogram. Reference-free alignment was performed in PyTom⁶⁸. The obtained volume was used as the template for a 6D correlation search of the same tomogram⁶⁹, yielding 675 subvolumes. Subvolumes were aligned and averaged in PyTom to obtain a 3D reconstruction with an estimated resolution of 28 Å (Fourier shell correlation, FSC = 0.5) (Extended Data Fig. 9a).

For negative-stain single-particle analysis, 21,365 particles were selected semi-automatically using *e2boxer.py* from EMAN2⁶⁷. All 3D reconstructions were performed in RELION⁷⁰. The unbiased negative-stain tomography reconstruction was low-pass filtered to 60 Å and used as an initial model for 3D reconstruction. The selected particles were extracted with a 160² pixel box and pre-processed to normalize images and to remove pixel values more than 5 standard deviations from the mean value. Particle images were sorted by unsupervised 3D classification in RELION⁷⁰ into four classes with the regularization parameter *T* set to 1.5, an initial angular sampling interval of 7.5°, an offset search range of 5 pixels, and offset search steps of 1 pixel (Extended Data Fig. 8a). This yielded one class of 8,815 particles that showed density for all components of the cITC–cMed. This class was refined using the 3D auto-refine procedure in RELION⁷⁰ with default parameters, to an estimated resolution of 25.2 Å with the ‘gold-standard’ FSC = 0.143 (Extended Data Fig. 9c–e). For validation of the reconstruction quality, reference-free 2D class averages were calculated from all particles included in the final reconstruction using RELION⁷⁰. The obtained averages were compared with SPIDER-generated⁷¹ forward projections of the final reconstruction (Extended Data Fig. 9b).

For cryo-EM single-particle analysis, 89,769 particles were selected semi-automatically using *e2boxer.py* from EMAN2⁶⁷. CTF parameters were estimated using CTFFIND⁷². CTF correction and 3D reconstruction were performed in RELION⁷⁰. The selected particles were extracted with a 280² pixel box and pre-processed to normalize images and remove pixel values more than 5 standard deviations from the mean value. Sorting of particle images by unsupervised 3D classification led to three classes, Pol II–DNA/RNA (14,777 particles), cITC (4,439 particles), and cITC–cMed (3,267 particles) (Extended Data Fig. 8b). Each class was refined only against the respective particles within this class using RELION⁷⁰. All reference maps were filtered to 60 Å before refinement using the 3D auto-refine procedure in RELION⁷⁰. Pol II–DNA/RNA and cITC classes were refined using a soft spherical reference mask (300 Å diameter), an initial angular sampling interval of 7.5°, an offset search range of 3 pixels, and offset search steps of 1 pixel. Pol II–DNA/RNA and cITC 3D reconstruction were refined to an estimated resolution of 6.6 Å and 7.8 Å, respectively, at FSC = 0.143 (Extended Data Fig. 9h, m). For visualization and structural modelling of the EM densities, temperature factors of –240 Å² (Pol II–DNA/RNA) and –340 Å² (cITC) were applied. The cITC–cMed class was refined using the 3D auto-refine procedure in RELION as Pol II–DNA/RNA and cITC, except with a soft SPIDER-generated⁷¹ reference mask in the shape of the cITC–cMed (maximum diameter of 360 Å) and an initial offset search range of 5 pixels. The cITC–cMed 3D reconstruction was refined to an estimated resolution of 9.7 Å at FSC = 0.143 (Extended Data Fig. 9r). For visualization and structural modelling of the cITC–cMed EM density, a temperature factor of –340 Å² was applied.

For validation of the reconstruction quality, reference-free 2D class averages were calculated from all particles included in the final reconstructions. The obtained averages were compared with SPIDER-generated⁷¹ forward projections of the final reconstructions (Extended Data Fig. 9f, k, p). To confirm the presence of only a single cITC–cMed conformation, we repeated 3D classification with rejected particles from class 1, 2, 5, 7, 8 of round 1 and cITC–cMed particles from class 4 of round 1 (see 3D classification of cryo-EM data) with the Pol II–DNA/RNA reconstruction filtered to 60 Å resolution as initial model (Extended Data Fig. 8c, d). For these 3D classifications, an initial angular sampling of 7.5°, an offset search range of 5 pixels, and offset search steps of 1 pixel were employed.

3D classification of cryo-EM data. Unsupervised 3D classification of the cryo-EM data was performed using RELION⁷⁰ in a pseudo-hierarchical manner that consisted of seven rounds of classification (Extended Data Fig. 8b). The obtained classes were iteratively reclassified and/or combined, based on the structures of Pol II⁴⁷, human minimal pre-initiation complex⁷ and cITC–cMed. The regularization parameter *T* was set to 4 for all rounds of classification. The negative-stain single-particle reconstruction of the cITC–cMed complex was low-pass filtered to 60 Å as the initial model for round 1 of 3D classification. The cryo-EM reconstruction of the human minimal pre-initiation complex (EMD-2305) was low-pass filtered to 60 Å as the initial model for round 2b of 3D classification. All classification was performed with soft reference masks as detailed for 3D refinement of Pol II–DNA/RNA, cITC and cITC–cMed (see Image processing).

Round 1 served to sort all 89,769 particles into eight classes to discard faulty particle images and identify a population of the complete cITC–cMed. The following initial sampling parameters were used: angular sampling of 7.5°, an offset search range of 15 pixels, and offset search steps of 3 pixels. Round 1 converged after 55 iterations and gave rise to Class 3 (~13,400 particles, 14.8% of data) that showed strong density for the cITC and partial density for cMed. This class was submitted to a second round of classification (round 2a) into four classes with an initial angular sampling of 7.5°, an offset search range of 5 pixels, and offset search steps of 1 pixel. All following rounds of classification were performed using the same initial sampling parameters used for round 2a. Class 3 of round 2a (3,267

particles, 3.7%) revealed the cITC–cMed complex with equal intensity of cITC and cMed, and was consequently refined as detailed (see Image processing). To explore any remaining heterogeneity, class 3 of round 2a was sorted into four classes (round 3a). This yielded class 1 and 2, lacking density for either upstream DNA–TFIIF–TFIIB–TBP or upstream DNA–TFIIB–TBP and the mobile plank of cMed (Extended Data Fig. 8b), and class 3 and 4 that varied appreciably in the mobile plank domain (Fig. 5d). Due to the small number of particle images, classes from round 3a were not refined.

Class 3 and 6 of round 1 showed density for Pol II and were consequently combined (~50,000 particles, 55.5%) and classified into eight classes (round 2b). This yielded class 2 (~14,000 particles, 15.5%) that presented a cITC with weak density for general transcription factors and upstream DNA. Class 2 of round 2b was therefore further classified into three classes (round 3b) that resulted in the cITC containing class 2 (~3,700 particles, 4.1%). To obtain a larger set of homogeneous cITC particles, class 1 of round 2a and class 2 of round 3b were combined (~7,300 particles, 8.1%) and sorted into four classes (round 4b). This led to class 2 of round 4b (4,439 particles, 4.9%) that revealed the complete cITC. The complete cITC was subsequently refined as detailed (see Image processing).

Class 1 and 8 of round 2b and class 3 of round 3b displayed density only for Pol II–DNA/RNA and were consequently combined (~25,000 particles, 27.7%) and further sorted into four classes (round 5). The resulting class 3 (14,777 particles, 16.4%) showed density for the Pol II–DNA/RNA complex with weak density for the Rpb4–Rpb7 subcomplex and was subsequently refined as detailed (see Image processing). Class 1 of round 5 (~3,200 particles, 3.5%) presented density for 12-subunit Pol II–DNA/RNA, but was not refined due to the small number of particles and poor orientational distribution (Extended Data Fig. 8b). Class 4 of round 2a and class 1 of round 3b displayed density for the binary Pol II–TFIIF–DNA/RNA complex and were combined (~12,900 particles, 15%, round 6). This complex presented with identical features found in the cITC and was not analysed further (Extended Data Fig. 8b).

Local resolution, filtering, and variance estimation. Local resolution maps (LRMs) were obtained using a method⁷³ implemented to run on a Graphics Processing Unit. A sliding window of 40³ voxels was centred around each voxel, extending the original half-maps through mirroring at the borders. The FSC at 0.143 was then calculated within the window and assigned to the central voxel. Given the low particle counts involved in the reconstructions and the resultant non-uniform sampling, five differently randomized pairs of half-maps generated in RELION⁷⁰ were processed independently and the results averaged to obtain a more robust estimate. Nevertheless, the maximal value for the local resolution was capped at the global FSC = 0.143 value to prevent enhancing of noise. Therefore, even though some regions, notably the Pol II density, clearly exhibited higher resolution than the nominal, LRMs were used only to limit the resolution locally, for a conservative interpretation of our data.

To perform local filtering, maps were downsampled for the Nyquist frequency to match the highest frequency value within the respective LRM. Look-up maps were then created by low-pass filtering the original map to each integer frequency value present in the LRM. For each voxel, a value linearly interpolated between the two look-up maps closest to its non-integer frequency was copied to the output. Finally, the locally filtered maps were rescaled to their original size.

Variance maps were generated for each cryo-EM reconstruction as an additional metric for assessing reconstruction quality and structural variability. These were obtained by implementing a described bootstrap approach⁷⁴, using the direct Fourier inversion method⁷⁵ for the reconstructions. As RELION’s data processing pipeline could not be used for this task, the aligned micrographs were CTF-corrected by means of automatic defocus determination and phase flipping in the TOM toolbox⁶⁴. Single particle views were then extracted using RELION’s position and rotation estimates. Differently sampled 12.5% fractions of the entire particle set were used to create 4,000 reconstructions. The variance between these reconstructions was calculated for each voxel position and normalized by the respective intensity value (Extended Data Fig. 9j, o, t).

Structural modelling. To generate unbiased models, known structures or homology models were sequentially rigid-body fitted using an automated global 6D correlation search in Situs⁷⁶ (Extended Data Figs 2a, 3a). As the majority of fitted models account for only a small fraction of the density, a Laplacian filter was applied for fitting (except for fitting the 10-subunit Pol II structure into the Pol II–DNA/RNA and cITC maps, and for fitting the cITC model lacking Rpb4–Rpb7 into the cITC–cMed map). After fitting of each component, difference maps were generated using UCSF Chimera⁷⁷ to reduce the search space in subsequent searches.

For the cITC model, we automatically fitted crystal structures of Pol II^{67,78}, TFIIB residues 22–213⁶, TBP⁵⁶, and homology models of TFIIF⁹, and the TFIIB C-terminal cyclin domain⁹ (Extended Data Fig. 2a). Pol II, DNA, and the Tfg1 charged helix were slightly adjusted manually using COOT⁷⁹ to accurately reflect the density. The crystal structure of the initially transcribing complex⁶ (PDB code

4A3D) was used for fitting of Pol II–DNA/RNA and initial fitting of cITC, as it presents the most complete model of Pol II except for the protrusion domain. For the Pol II–DNA/RNA model Rpb4–Rpb7 was excluded, due to the weak density in this region. For fitting of the 10-subunit Pol II–TFIIB model, we combined structures of initially transcribing Pol II (PDB code 4A3D, excluding Rpb4–Rpb7) and initially transcribing Pol II–TFIIB complex (PDB code 4BBS, chain M, TFIIB residues 22–118). After visual inspection of the cITC density, the nucleic acids and Pol II domains clamp core, wall and protrusion in the initially transcribing Pol II structure (PDB code 4A3D) were replaced with the corresponding regions from the initially transcribing Pol II–TFIIB complex structure (PDB code 4BBS). For fitting of Rpb4–Rpb7, chains D and G from the initially transcribing Pol II structure (PDB code 4A3D) were used.

For fitting the structure of the TFIIB N-terminal cyclin domain–TBP–DNA complex, the atomic coordinates of the equivalent *Homo sapiens* complex (PDB code 1VOL) were used as a template. TBP and the TFIIB N-terminal cyclin of 1VOL were replaced with *S. cerevisiae* TBP (PDB code 1YTB, chain A) and the *S. cerevisiae* TFIIB N-terminal cyclin (PDB code 4BBS, chain M and residues 122–213), respectively. The model of the TFIIF dimerization domain was generated as described⁹, containing *S. mikatae* Tfg1 residues 92–143 and 343–412, and *S. cerevisiae* Tfg2 residues 54–138 and 208–227. The homology model of the TFIIB C-terminal cyclin (residues 233–342) was based on the *H. sapiens* TFIIB C-terminal cyclin⁹ (PDB code 1VOL). The Tfg1 charged helix was placed manually using COOT⁷⁹ according to the chemical environment. Further density in shape of duplex DNA was observed upstream and downstream of TATA-box DNA, and downstream of the active site. This density was modelled with canonical duplex B-DNA, generated in COOT⁷⁹. NAM⁸⁰ was used to optimize DNA geometry using the CHARMM force field⁸¹. The DNA sequence of all nucleic acid models was edited in COOT⁷⁹ to match the sequence of the employed nucleic acid scaffold.

The cITC–cMed model was generated by sequential fitting of cITC lacking Rpb4–Rpb7, and the three regions of an improved *S. cerevisiae* Mediator head module model (Extended Data Fig. 3a). The construction of an improved head module was required because the published structures are at 4.3 Å resolution and contain out-of-register errors and are lacking several regions^{13,15}. To account for the known flexibility within the head module¹⁴, its regions neck (including joint), movable jaw, and fixed jaw were fitted independently. To model the fixed jaw tooth and the nose domains, we used the structure of *S. cerevisiae* Med17C–Med11C–Med22C (PDB code 4H62¹⁴). For the movable jaw, the structure of *S. cerevisiae* Med8C–Med18–Med20 (PDB code 2HZS³²) was completed with Med18 residues 69–93 from the structure of *S. cerevisiae* Med18–Med20 (PDB code 2HZM, chain D) to account for additional residues visible in the electron density. The shoulder domain was modelled with MODELLER⁸² based on its *S. pombe* counterpart (PDB code 4H61¹⁴). For the arm, finger, spine and joint domains of the head module, our *S. pombe* head module (PDB code 4H63¹⁴) was used as a model. Med11 and Med22 helices $\alpha 1$ and $\alpha 2$ were replaced with the corresponding helices from the *S. cerevisiae* Med11N–Med22N structure (PDB code 3R84). Other *S. pombe* residues were replaced with their *S. cerevisiae* counterparts, based on sequence alignments. The obtained models for all head module domains except the finger were individually fitted onto the corresponding elements in the published structure of the *S. cerevisiae* head module (PDB code 4GWP), using secondary structure matching in COOT⁷⁹. A 2F_o–F_c electron density map corresponding to the published *S. cerevisiae* head module structure was calculated. The map was used to fit our model of the finger element and to adjust the rest of our model when required. The linker between helices $\alpha 4$ and $\alpha 5$ in Med8 and the *Sc* specific helix between strand $\beta 1$ and helix $\alpha 3$ were built manually. The geometry of the obtained head module model was regularized in COOT⁷⁹ and with PHENIX⁸³. All figures and movies were generated using UCSF Chimera⁷⁷ and PyMol.

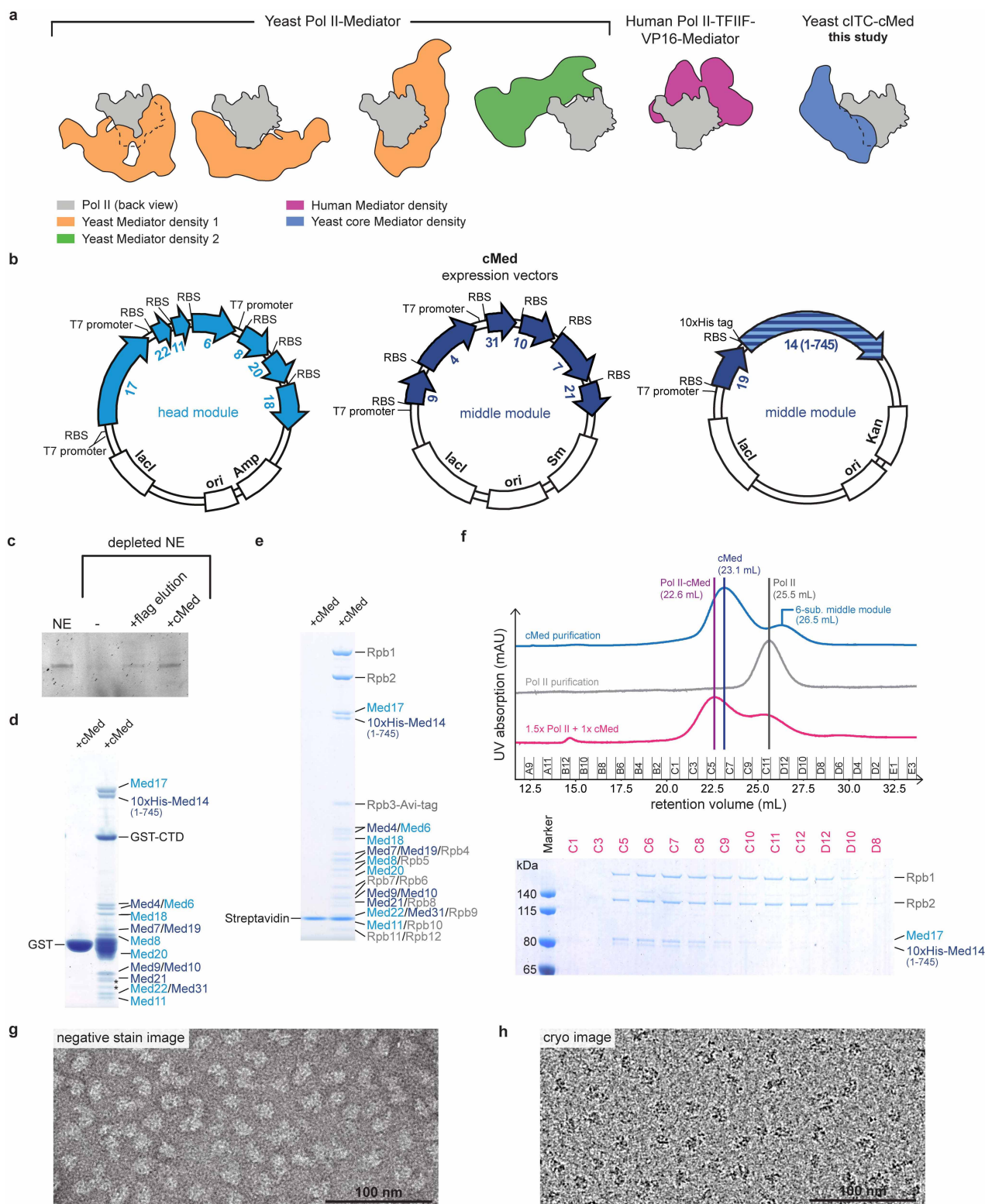
Immobilized template assay. The immobilized template assay was performed essentially as described⁵³ using nuclear extract from *srb4-ts*⁴² or wild-type (BY4741) *S. cerevisiae* strains and a linear *HIS4* promoter^{53,54}. The *srb4-ts* extract was complemented with recombinant head module (4.5 pmol) as indicated in Extended Data Fig. 5b. Samples were applied to SDS–PAGE, transferred onto a polyvinylidene difluoride membrane (Millipore) and probed by antibodies anti-Rpb3 (1Y26, NeoClone, dilution 1:1000), anti-TFIIB (ab63909, Abcam, dilution 1:4000), anti-TBP (sc-33736, Santa Cruz, dilution 1:400), anti-Med17 (1:10000 dilution), kindly provided by Steven Hahn (Fred Hutchinson Cancer Research Center), anti-Flag-tag (F1804, Sigma, 1:1000 dilution), and anti-Med2 (sc-28058, Santa Cruz, 1:1000 dilution). Secondary antibodies anti-rat IgG HRP (A9037, Sigma, dilution 1:3000), anti-mouse IgG HRP (170-6516, Bio-Rad, dilution 1:3000), anti-rabbit IgG HRP (sc-2004, Santa Cruz, dilution 1:3000) and anti-goat IgG HRP (sc-2020, Santa Cruz, dilution 1:3000) were used. Antibody detection was achieved with Pierce enhanced chemiluminescence (ECL) western blotting substrate (Thermo Scientific) and Amersham Hyperfilm ECL (GE Healthcare).

Comparative Dynamic Transcriptome Analysis (cDTA). cDTA enables global analysis of newly synthesized RNA⁴⁴ that reveals defects in transcription with much higher sensitivity than conventional steady-state methods. cDTA was carried out as described⁴⁴ using the *S. cerevisiae* heterozygous *Med17/med17Δ* strain (Euroscarf) transfected with plasmids pRS315-SRB4¹⁵ or pRS315-srb4-ts¹⁵, and Y40343-wildtype (Euroscarf) or Med18-FRB-KanMX6 (Euroscarf) strains. Heat shock of SRB4 and *srb4-ts* strains was applied for 18 or 60 min at 37 °C before RNA labelling as described⁴⁴. To deplete the Med18 subunit from the nucleus, anchor-away experiments were performed by rapamycin treatment (1 μg ml^{−1} in 200 ml YPD) for 18 or 60 min at 30 °C before RNA labelling⁸⁴. Data analysis was as described⁴⁴.

CTD phosphorylation assay. Endogenous TFIIF was purified as described³¹. Purified Pol II (2.5 pmol), 7-subunit head module (1.25 pmol), 6-subunit middle module (1.25 pmol), cMed (2.5 pmol) were added as indicated in Extended data Fig. 7f and incubated on ice for 10 min before addition of TFIIF (~0.02 pmol) and 300 μM ATP. Reactions were incubated for 25 min at 25 °C, applied to SDS–PAGE, transferred to a nitrocellulose membrane (GE Healthcare) and probed with primary antibodies anti-Ser5-P (3E10, dilution 1:20), provided by D. Eick (Helmholtz-Zentrum München), and anti-Rpb3 (1Y26, NeoClone, dilution 1:2,000). Secondary antibodies anti-rat IgG HRP (A9037, Sigma, dilution 1:3,000) and anti-mouse IgG HRP (170-6516, Bio-Rad, dilution 1:3,000) were used. Antibody detection was achieved with Pierce enhanced chemiluminescence (ECL) western blotting substrate (Thermo Scientific) and an Advanced Fluorescence Imager (Intas).

51. Li, Y. *et al.* Yeast global transcriptional regulators Sin4 and Rgr1 are components of mediator complex/RNA polymerase II holoenzyme. *Proc. Natl Acad. Sci. USA* **92**, 10864–10868 (1995).
52. Baumli, S., Hoepfner, S. & Cramer, P. A conserved mediator hinge revealed in the structure of the MED7.MED21 (Med7.Srb7) heterodimer. *J. Biol. Chem.* **280**, 18171–18178 (2005).
53. Seizl, M., Larivière, L., Pfaffeneder, T., Wenzek, L. & Cramer, P. Mediator head subcomplex Med11/22 contains a common helix bundle building block with a specific function in transcription initiation complex stabilization. *Nucleic Acids Res.* **39**, 6291–6304 (2011).
54. Rani, P. G., Ranish, J. & Hahn, S. RNA polymerase II (Pol II)–TFIIF and Pol II–mediator complexes: the major stable Pol II complexes and their activity in transcription initiation and reinitiation. *Mol. Cell. Biol.* **24**, 1709–1720 (2004).
55. Sydow, J. F. *et al.* Structural basis of transcription: mismatch-specific fidelity mechanisms and paused RNA polymerase II with frayed RNA. *Mol. Cell* **34**, 710–721 (2009).
56. Kim, Y., Geiger, J. H., Hahn, S. & Sigler, P. B. Crystal structure of a yeast TBP/TATA-box complex. *Nature* **365**, 512–520 (1993).
57. Chen, H.-T., Warfield, L. & Hahn, S. The positions of TFIIF and TFIIE in the RNA polymerase II transcription preinitiation complex. *Nature Struct. Mol. Biol.* **14**, 696–703 (2007).
58. Kireeva, M. L., Lubkowska, L., Komissarova, N. & Kashlev, M. Assays and affinity purification of biotinylated and nonbiotinylated forms of double-tagged core RNA polymerase II from *Saccharomyces cerevisiae*. *Methods Enzymol.* **370**, 138–155 (2003).
59. Leitner, A. *et al.* Probing native protein structures by chemical cross-linking, mass spectrometry, and bioinformatics. *Mol. Cell. Proteomics* **9**, 1634–1649 (2010).
60. Herzog, F. *et al.* Structural probing of a protein phosphatase 2A network by chemical cross-linking and mass spectrometry. *Science* **337**, 1348–1352 (2012).
61. Leitner, A., Walzthoen, T. & Aebersold, R. Lysine-specific chemical cross-linking of protein complexes and identification of cross-linking sites using LC-MS/MS and the xQuest/xProphet software pipeline. *Nature Protocols* **9**, 120–137 (2014).
62. Petrochenko, E. V. & Borchers, C. ICC-CLASS: isotopically-coded cleavable crosslinking analysis software suite. *BMC Bioinformatics* **11**, 64 (2010).
63. Mastronarde, D. N. Automated electron microscopy tomography using robust prediction of specimen movements. *J. Struct. Biol.* **152**, 36–51 (2005).
64. Korinek, A., Beck, F., Baumeister, W., Nickell, S. & Plitzko, J. M. Computer controlled cryo-electron microscopy–TOM² a software package for high-throughput applications. *J. Struct. Biol.* **175**, 394–405 (2011).
65. Li, X. *et al.* Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nature Methods* **10**, 584–590 (2013).
66. Eibauer, M. *et al.* Unraveling the structure of membrane proteins in situ by transfer function corrected cryo-electron tomography. *J. Struct. Biol.* **180**, 488–496 (2012).
67. Tang, G. *et al.* EMAN2: an extensible image processing suite for electron microscopy. *J. Struct. Biol.* **157**, 38–46 (2007).
68. Chen, Y., Pfeffer, S., Hrabe, T., Schuller, J. & Förster, F. Fast and accurate reference-free alignment of subtomograms. *J. Struct. Biol.* **182**, 235–245 (2013).
69. Hrabe, T. *et al.* PyTom: a python-based toolbox for localization of macromolecules in cryo-electron tomograms and subtomogram analysis. *J. Struct. Biol.* **178**, 177–188 (2012).
70. Scheres, S. H. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180**, 519–530 (2012).
71. Shaikh, T. R. *et al.* SPIDER image processing for single-particle reconstruction of biological macromolecules from electron micrographs. *Nature Protocols* **3**, 1941–1974 (2008).
72. Mindell, J. A. & Grigorieff, N. Accurate determination of local defocus and specimen tilt in electron microscopy. *J. Struct. Biol.* **142**, 334–347 (2003).

73. Cardone, G., Heymann, J. B. & Steven, A. C. One number does not fit all: mapping local variations in resolution in cryo-EM reconstructions. *J. Struct. Biol.* **184**, 226–236 (2013).
74. Penczek, P. A., Yang, C., Frank, J. & Spahn, C. M. T. Estimation of variance in single-particle reconstruction using the bootstrap technique. *J. Struct. Biol.* **154**, 168–183 (2006).
75. Penczek, P. A., Renka, R. & Schomberg, H. Gridding-based direct Fourier inversion of the three-dimensional ray transform. *J. Opt. Soc. Am. A* **21**, 499–509 (2004).
76. Wriggers, W. Using Situs for the integration of multi-resolution structures. *Biophys. Rev.* **2**, 21–27 (2010).
77. Pettersen, E. *et al.* UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
78. Cheung, A. C. M., Sainsbury, S. & Cramer, P. Structural basis of initial RNA polymerase II transcription. *EMBO J.* **30**, 4755–4763 (2011).
79. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
80. Phillips, J. C. *et al.* Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **26**, 1781–1802 (2005).
81. Brooks, B. R. *et al.* CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **4**, 187–217 (1983).
82. Eswar, N. *et al.* Comparative protein structure modeling using Modeller. *Curr. Protocols Bioinf.* **5**, 5.6.1–5.6.30 (2006).
83. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
84. Sun, M. *et al.* Global analysis of eukaryotic mRNA degradation reveals Xrn1-dependent buffering of transcript levels. *Mol. Cell* **52**, 52–62 (2013).
85. Cai, G., Imasaki, T., Takagi, Y. & Asturias, F. J. Mediator structural conservation and implications for the regulation mechanism. *Structure* **17**, 559–567 (2009).
86. Cai, G. *et al.* Interaction of the mediator head module with RNA polymerase II. *Structure* **20**, 899–910 (2012).
87. Koschubs, T. *et al.* Preparation and topology of the Mediator middle module. *Nucleic Acids Res.* **38**, 3186–3195 (2010).



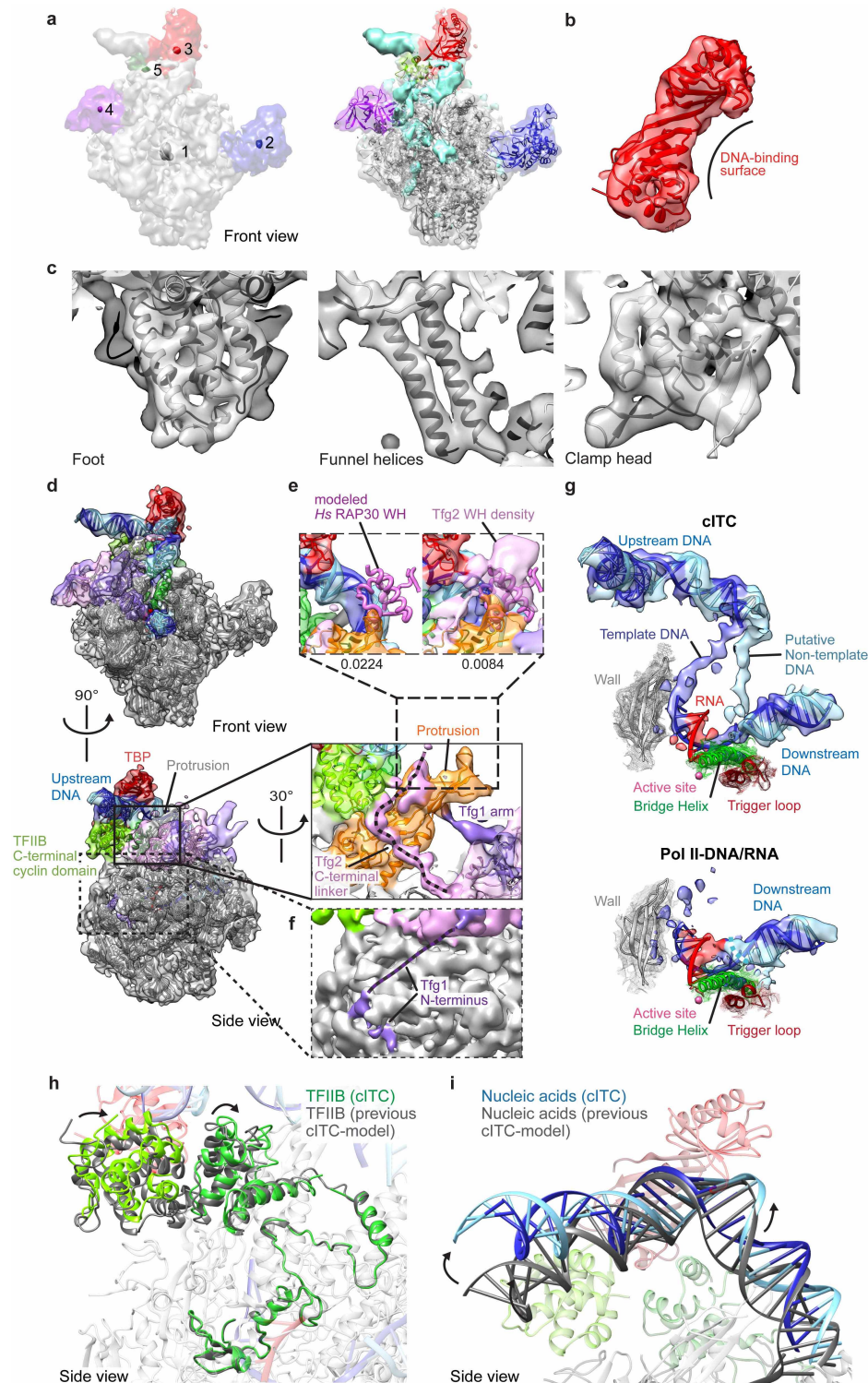
Extended Data Figure 1 | cMed reconstitution, activities, and cITC binding.

a, Cartoon view of the different reported positions of Mediator on Pol II. Yeast Mediator density was observed at four different locations on Pol II (orange left¹⁹, middle^{13,85,86}, right¹⁵, green²²). Another Mediator position was obtained in the human Pol II-TFIIF-VP16-Mediator complex²¹ (purple). The Mediator position presented in this work is shown for comparison (blue). Pol II is oriented the same way in all views, approximately viewed from the back.

b, Schematic view of vectors used for co-expression of cMed. RBS, ribosome binding site; 10xHis tag, 10xhistidine tag; *ori*, origin of replication; *lacI*, gene encoding Lac repressor. *Sm*, *Amp* and *Kan* refer to streptomycin, ampicillin and kanamycin resistance genes, respectively.

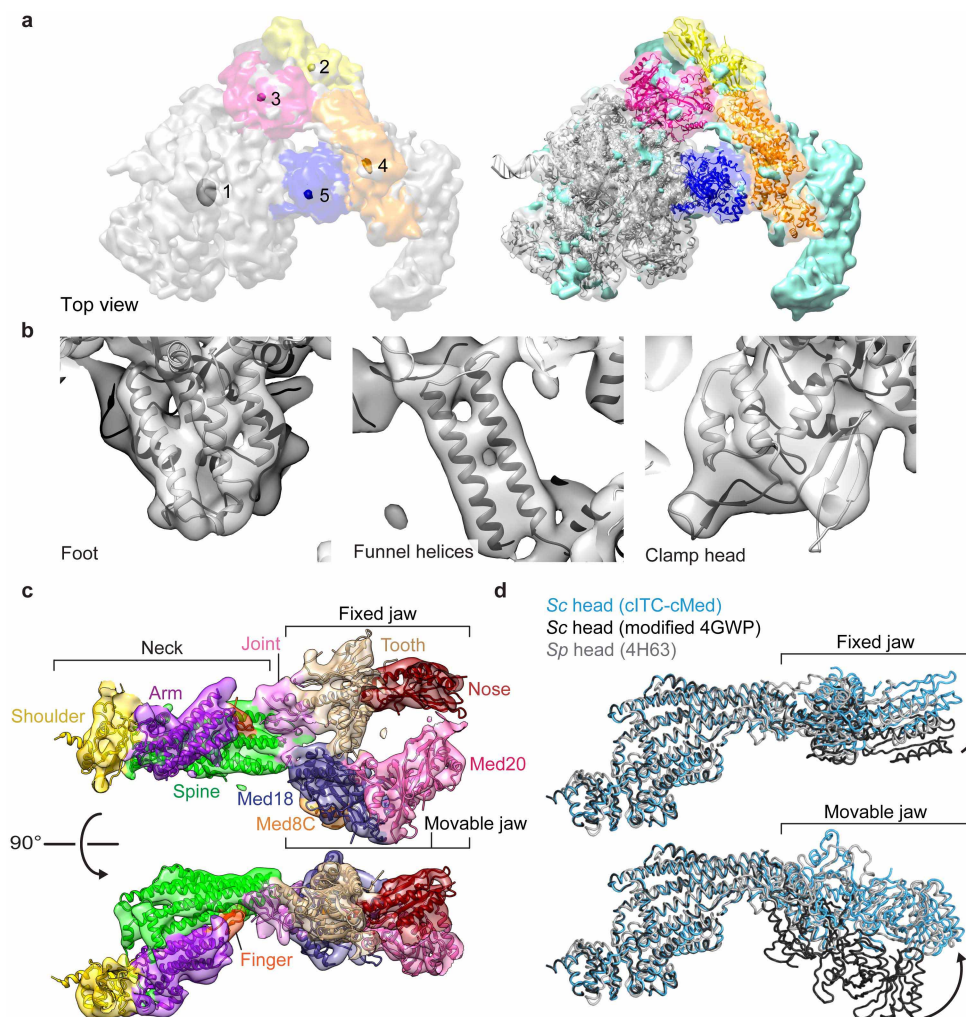
c, Recombinant cMed is active in

activator-independent promoter-dependent transcription⁵³. Compare with Fig. 1b. **d**, Recombinant cMed binds purified GST-CTD fusion protein, but not GST alone that was immobilized on glutathione resin. An asterisk marks two contaminant bands. **e**, Recombinant cMed binds Pol II in a pull-down experiment. Pol II was immobilized on streptavidin resin via a biotin tag on the Rpb3 subunit. **f**, Recombinant cMed binds purified Pol II during size exclusion chromatography. Chromatograms of cMed (blue), Pol II (grey) and Pol II-cMed complex (magenta) are shown (top). Peak fractions of the Pol II-cMed were analysed by SDS-PAGE (bottom). **g**, Negative-stain EM image of cITC-cMed complex. The scale bar is 100 nm. **h**, Cryo-EM image of cITC-cMed complex.



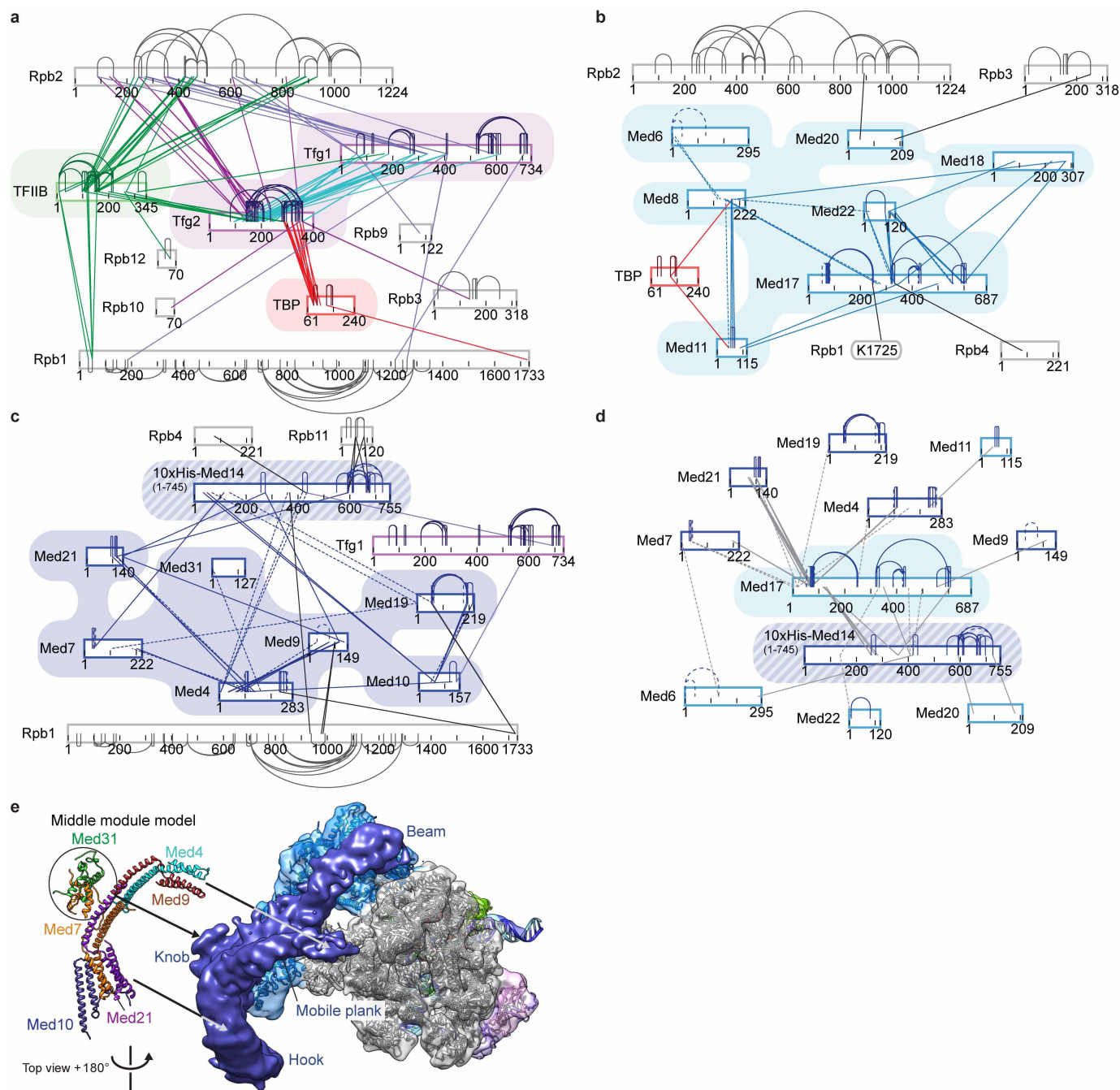
Extended Data Figure 2 | Structural modelling of cITC into the cryo-EM density. **a**, Automatic fitting of structures into cITC cryo-EM density. The order of structure fitting and the corresponding translation correlation peaks are indicated (left). After fitting of all structures (right), the remaining density (cyan) was attributed mainly to DNA and TFIIF. **b**, Fit of TBP to the cITC EM density. **c**, Detailed views of Pol II domains foot, funnel, and clamp head in the cITC EM density. **d**, Two views of the cITC EM reconstruction corresponding to previously defined front and side views of Pol II⁴⁷. The final cITC model is coloured (DNA template/non-template, blue/cyan; RNA, red; Pol II, silver; TBP, red; TFIIB, green; TFIIF Tfg1/Tfg2, violet/magenta). Tfg1 and Tfg2 contain non-conserved insertions in the TFIIF dimerization module. **e**, The cITC map reveals density for the Tfg2 C-terminal linker and winged helix (WH) domain. The view corresponds to the side view in **c**, but is

rotated around a vertical axis by 30°. The mobile Tfg2 WH domain is visible at lower density threshold. The homologous human RAP30 WH domain (*Homo sapiens*, Hs) was modelled on the basis of its position in the human initiation complex⁷ and resides at a similar location. **f**, The cITC map reveals density that may correspond to the Tfg1 N terminus, as suggested by protein-protein crosslinking^{9,30}. **g**, Comparison of EM densities for promoter DNA in cITC (top) and in the Pol II-DNA/RNA complex (bottom). The Pol II elements active site (magenta), bridge helix (green), trigger loop (dark red), wall (grey), and template DNA (dark blue), non-template DNA (light blue) and RNA (red) are depicted. **h**, Minor repositioning of TFIIB cyclin domains compared to our previous model⁹. **i**, Minor rotation of the TBP-DNA-TFIIB complex on the Pol II wall in the cITC structure compared to the previous open complex model⁹.



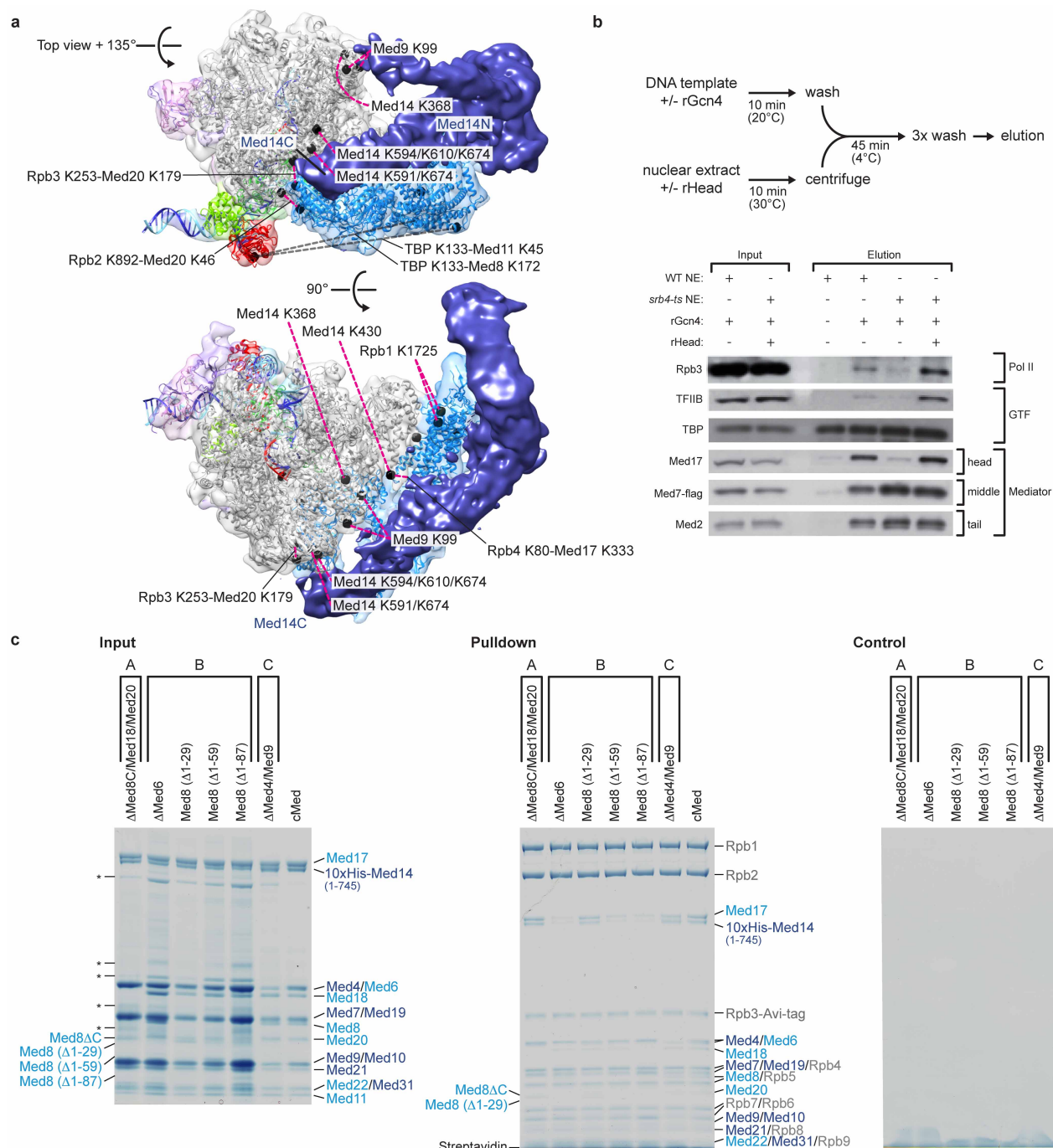
Extended Data Figure 3 | Structural modelling of cITC-cMed into the cryo-EM density. **a**, Automatic fitting of atomic models into the cITC-cMed density. The order of model fitting and the corresponding translation correlation peaks are indicated (left). After fitting of all models (right), the remaining density (cyan) was attributed to the middle module and to some minor additional protein regions in the cITC and head module. **b**, Detailed views of Pol II domains foot, funnel, and clamp head in the cITC-cMed EM density. **c**, Fit of improved Mediator head crystallographic model to the

corresponding cryo-EM density for cITC-cMed. The different head domains are depicted in different colours. Views are from the previously defined right side and top views of the head module¹⁴. **d**, The head module undergoes minor changes in conformation upon formation of the cITC-cMed complex. The EM fit with a modified model of the *S. cerevisiae* (Sc) Mediator head is compared to the crystal structures of head modules from *S. pombe* (Sp) (PDB code 4H63) and Sc (modified based on PDB code 4GWP). Previously defined top view of the head module¹⁴.



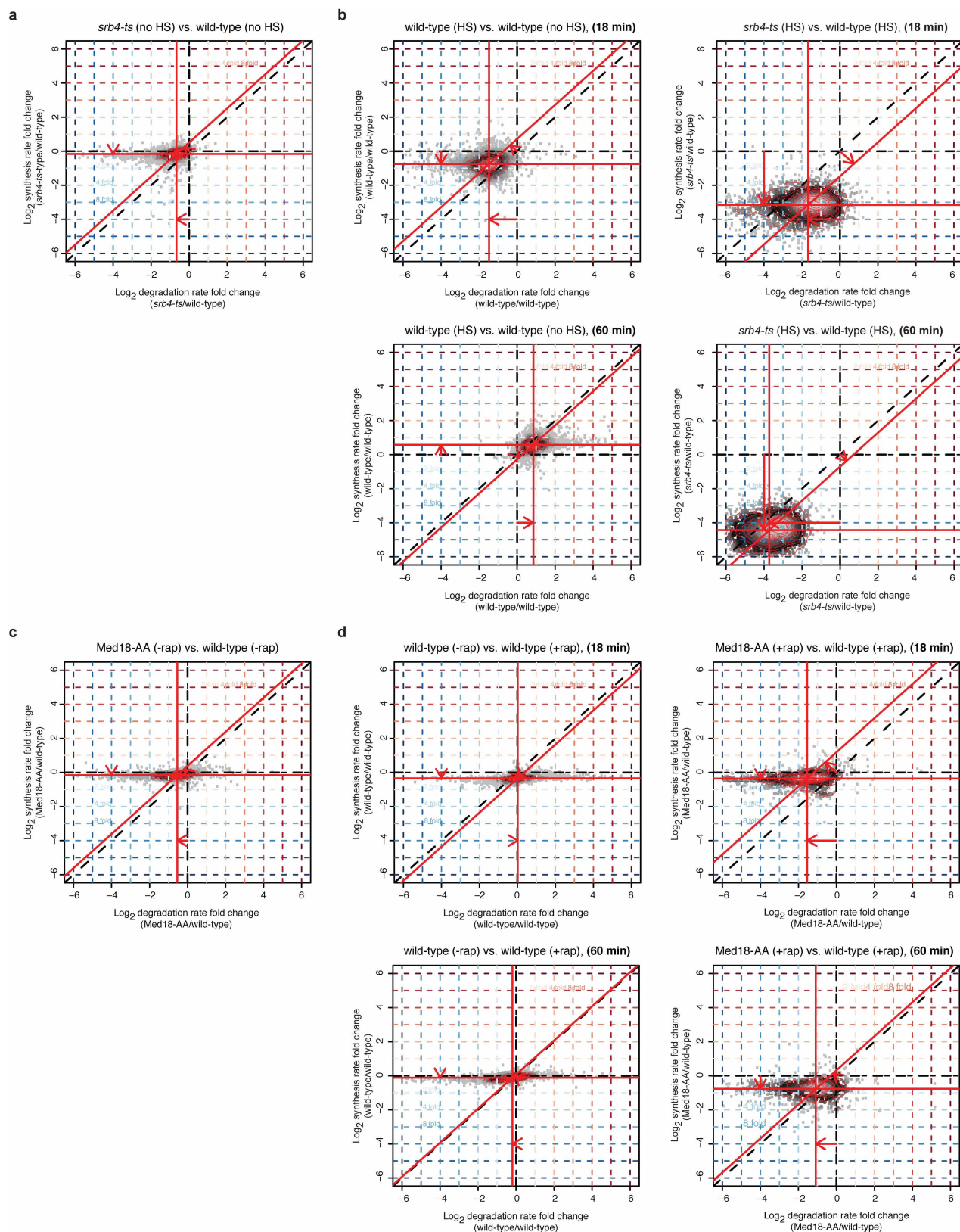
Extended Data Figure 4 | Protein-protein crosslinking. **a**, Crosslinks between Pol II and TBP, TFIIB, and TFIIF confirm cITC architecture. Crosslinks were visualized using xiNET (Rappsilber laboratory). **b**, Crosslinks within the cMed head module and between the head module and Pol II. Solid lines indicate crosslinks derived from cITC-cMed data; dashed lines indicate additional crosslinks obtained only for free cMed. **c**, Crosslinks within the cMed middle module and between the middle module and Pol II.

Crosslinks within the middle module agree well with the proposed middle module architecture^{25,87}. **d**, Crosslinks between head and middle modules elucidate cMed architecture. **e**, Possible location of the previous model of the Mediator middle module (ribbons, left²⁵) within the observed cMed density in the cITC-cMed reconstruction based on protein crosslinking. The cITC-cMed complex is viewed from the 'bottom', which corresponds to the top view rotated by 180° around a vertical axis.



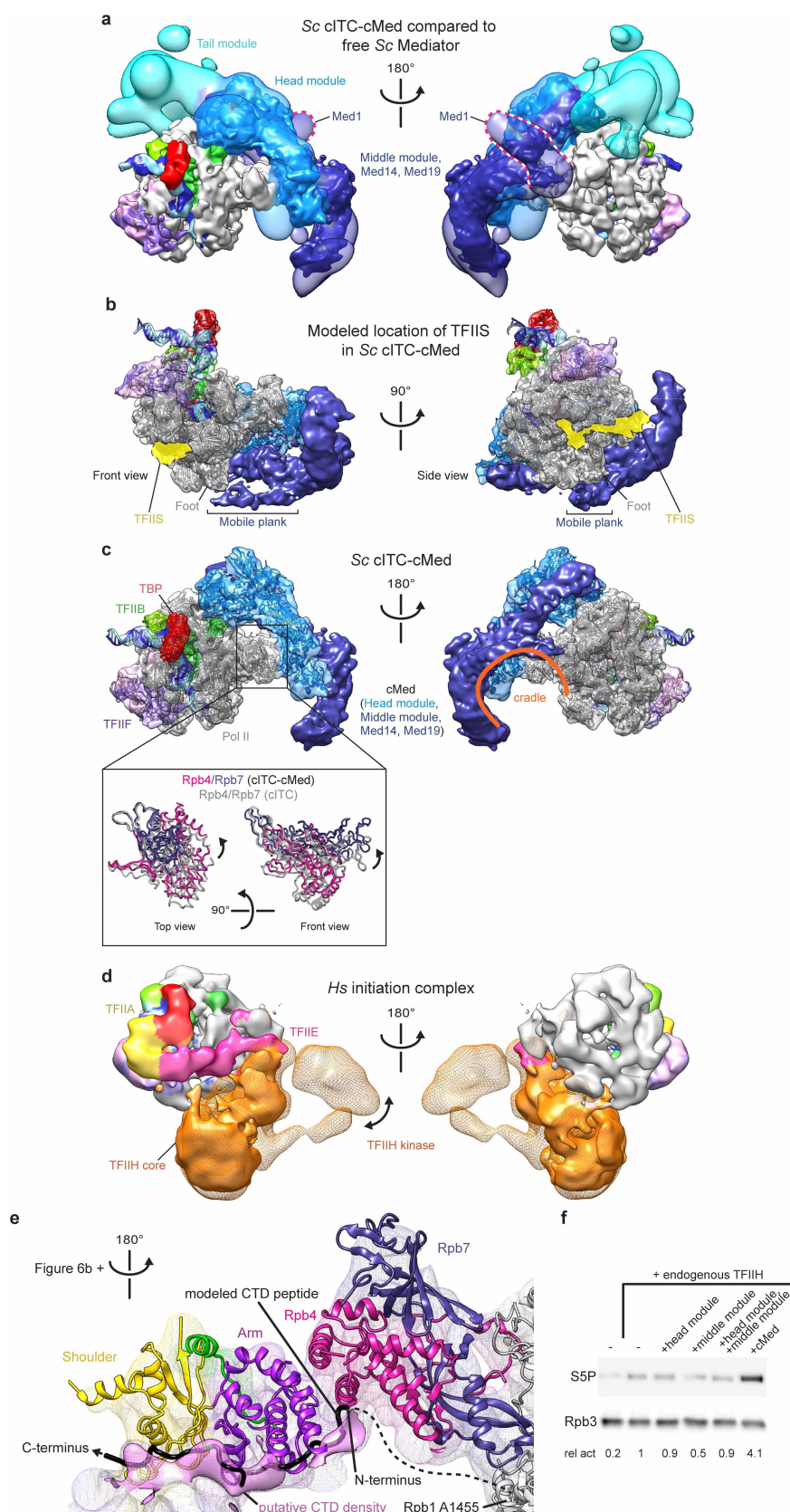
Extended Data Figure 5 | Analysis of the cITC-cMed interface. **a**, Crosslinks between cITC and cMed were mapped on available models of cITC and Mediator head module. Two views of cITC-cMed are shown, a top view rotated by 135° around the horizontal axis and an additional 90° around the same axis. Mapped crosslinks exceeding the 30 Å distance restraint were coloured in grey. Crosslinks between lysine residues (black spheres) are labelled or connect (dashed magenta lines) to proximal cMed density, where plausible. Med14 N- and C-terminal regions are indicated. **b**, Immobilized template assay using nuclear extract of *srb4-ts* and wild-type yeast strains demonstrates that recruitment of Pol II and TFIIB depends on the Mediator head module. Schematic protocol for the immobilized template (ITA) assay⁵³ (top). The loss of Pol II, TFIIB and Mediator head module from promoter DNA by heat inactivation of the *srb4-ts* nuclear extract is rescued by addition of recombinant head (rHead) module⁵³ (bottom). **c**, Pull-down assays with recombinant cMed variants carrying mutations in interfaces A, B, and C reveal that interface B is essential for Pol II binding, whereas interfaces A and C are not. For definition of interfaces see Fig. 5. The assay measures retention of cMed II coupled to beads (see Extended Data Fig. 1e). Interface A was perturbed by

deletion of the movable jaw, Med8C(residues 1–189)–Med18–Med20. Interface B was perturbed by removal of parts of the arm domain, either by N-terminal truncation of Med8 (Δ1–29, deleting only the flexible N-terminal tail; Δ1–59, deleting the tail and helix α1; Δ1–87, deleting the tail and helices α1 and α2), or by removal of Med6, which contributes a helix to the arm domain (ΔMed6). Interface C was perturbed by deletion of Med4–Med9. Shown are SDS-PAGE analyses (Coomassie staining). On the left, 2 μg of purified input cMed variants were analysed and the integrity of the complexes confirmed. The identity of the cMed variant-specific bands was confirmed by mass spectrometry. Some minor contaminant bands are marked with asterisks. On the middle gel, the bead elutions after the pull-down assays are shown. Wild-type cMed bound to Pol II-coupled beads, providing a positive control (rightmost lane). cMed variants were retained by Pol II-coupled beads to various degrees. Whereas cMed variants with perturbations in interfaces A and C bound to Pol II, several variants with perturbations in interface B were impaired for Pol II binding. On the right, the bead controls are shown. None of the analysed cMed variants bound to beads only, providing a negative control.



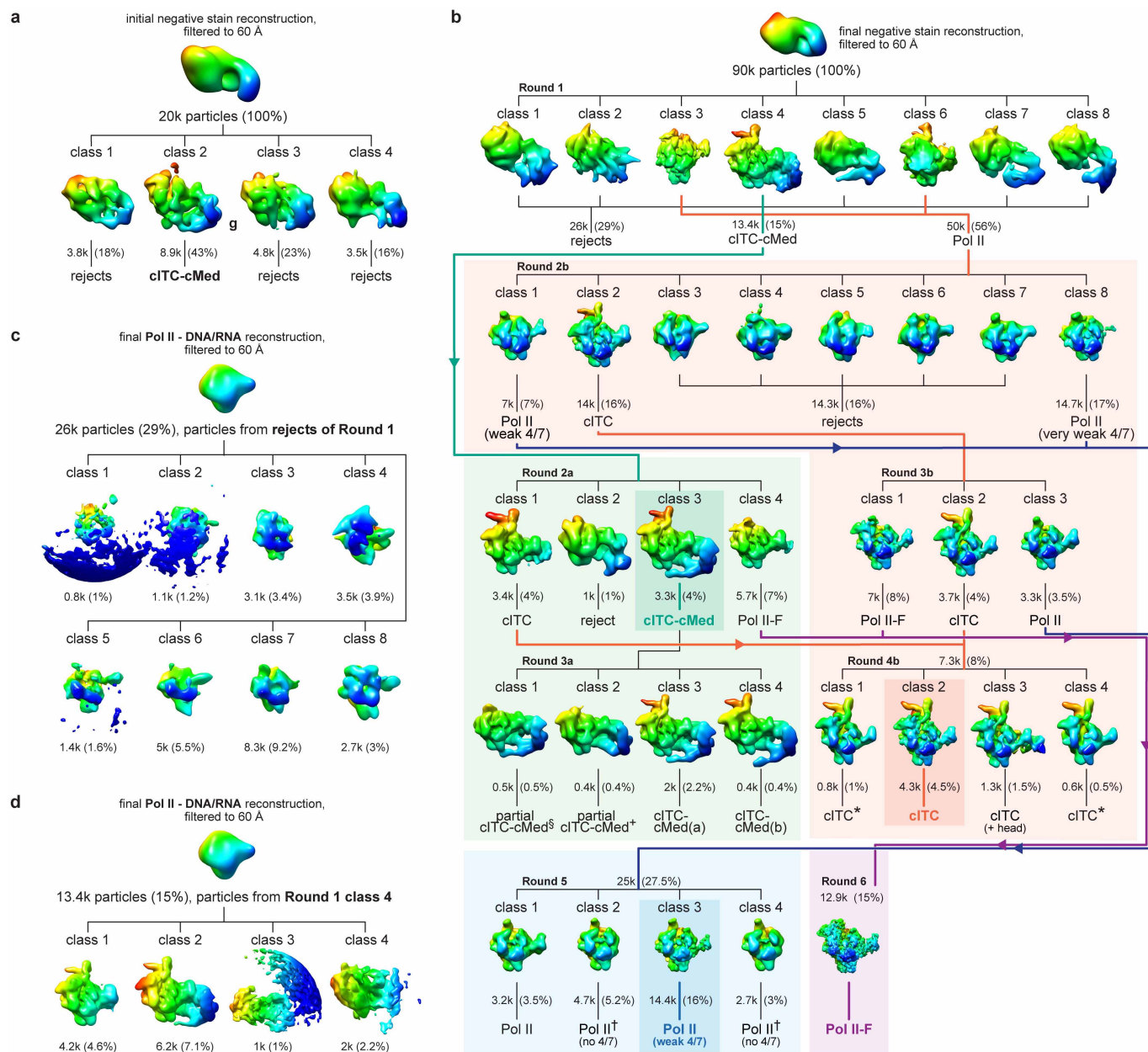
Extended Data Figure 6 | Global requirement of the Mediator head module for transcription. **a**, Fold changes in RNA degradation (log folds, *x*-axis) and synthesis (log folds, *y*-axis) rates observed in strains *srb4-ts* versus wild-type in the absence of heat shock. Each point corresponds to one mRNA and the density of points is reflected in their brightness. Red contour lines define regions of equal intensity. The centre of the distribution results from the median synthesis and degradation rates, whose relative contributions are indicated by shifts of the red lines parallel to synthesis or degradation rate axis,

respectively. **b**, Global shutdown of RNA synthesis upon heat shock (HS) of the *srb4-ts* mutant. Fold changes in degradation (log fold, *x*-axis) and synthesis (log folds, *y*-axis) rates of *srb4-ts* and wild-type strains are indicated, after 18 or 60 min of HS treatment, respectively. **c**, As for **a**, but using a Med18-anchor-away (AA) strain in absence of rapamycin (rap) treatment. **d**, Global downregulation of RNA synthesis upon anchor-away of the Med18 subunit, after 18 or 60 min of rapamycin (rap) treatment, as in **b**.



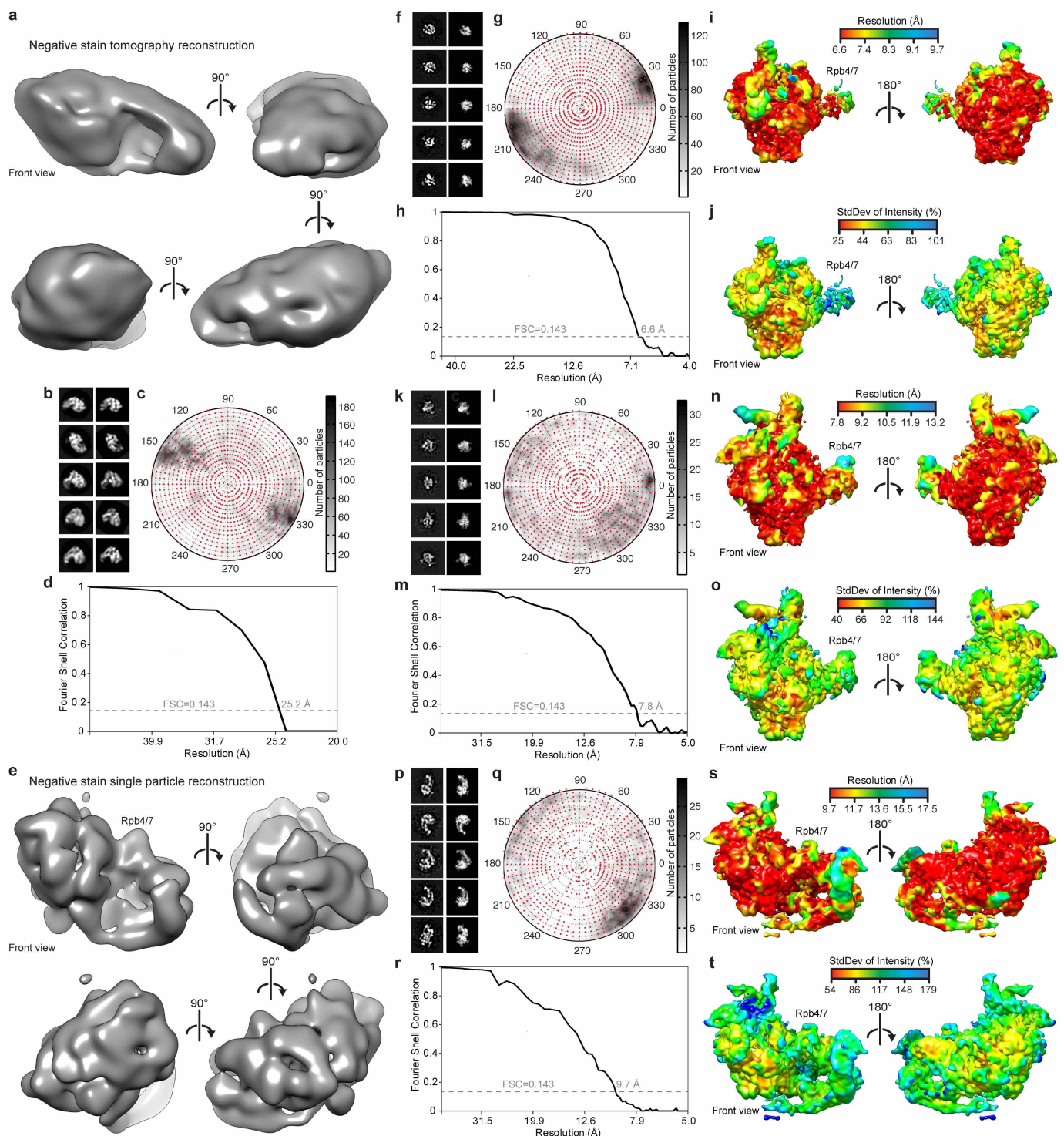
Extended Data Figure 7 | Inferred locations of the Mediator tail module, TFIIIS, the general factors TFIIA, TFIIIE, and TFIIH, and the CTD, and comparison with the human initiation complex. **a**, Superposition of revised free yeast Mediator EM reconstruction¹⁷ (transparent lower-resolution surface, coloured in shades of blue) onto cITC–cMed (coloured as in Fig. 3). This reveals a high degree of similarity in the cMed region, and suggests locations of the tail module and subunit Med1 that are not present in cMed. **b**, Modelled location of TFIIIS (yellow) within the cITC–cMed complex based on the Pol II–TFIIIS complex (PDB code 1Y1V)⁴⁸. **c**, Movement of the Rpb4–Rpb7 stalk upon cMed binding to the cITC and location of the ‘cradle’. Two views of the *S. cerevisiae* (Sc) cITC–cMed complex related by a 180° rotation around a vertical axis. The left view corresponds to the previously defined top view of Pol II⁴⁷. The box contains a zoom-in view of the Rpb4–Rpb7 complex revealing its movement upon cMed binding. The right view reveals the location of the ‘cradle’ on the ‘bottom’ of Pol II. The density is transparent, with the final cITC–head module model underneath and coloured (DNA template/non-template, blue/cyan; RNA, red; Pol II, silver; TBP, red; TFIIIB, green; TFIIIF Tfg1/Tfg2, violet/magenta; head module, blue; middle module, violet).

d, Human (*Homo sapiens*, Hs) initiation complex EM reconstruction⁷ viewed as in **a**. The complex reveals equivalent locations for TFIIIB, TBP, TFIIIF, and DNA, and additionally contains TFIIA (yellow), TFIIIE (magenta), and the core subcomplex of TFIIH (solid orange surface) and a modelled TFIIH kinase subcomplex (orange mesh). **e**, Putative density for the CTD (magenta) near a modelled peptide (black) positioned according to the head module–CTD complex co-crystal structure (PDB code 4GWQ). The putative density superposes moderately well with the modelled CTD peptide and suggests that the CTD may adopt a different conformation in cITC–cMed with respect to the binary head module–CTD complex. This density cannot be assigned to the CTD with certainty because it may also stem from unresolved protein regions in Mediator. Coloured as in Fig. 5a. The region C-terminal of the CTD remained flexible and crosslinked to distant proteins (not shown, see Extended Data Fig. 4). **f**, Recombinant cMed stimulates TFIIH kinase activity, whereas the free head and middle modules do not. The relative activity (rel act) of Pol II phosphorylation at CTD residue serine-5 (S5P) was determined with respect to background (Rpb3).



Extended Data Figure 8 | 3D classification of negative-stain and cryo-EM data. **a**, 3D classification of the negative-stain EM data set into four classes. The percentage of data in each class is given in parenthesis. To help visualize structural differences, all 3D reconstructions were radially coloured in UCSF Chimera. **b**, Pseudo-hierarchical 3D classification of the cryo-EM data set. The percentage of the data in each class is given in parentheses. Rejects refer to EM reconstructions that did not reflect the known structures of Pol II, cITC or cITC-cMed. §, class of partial cITC-cMed particles lacking upstream DNA-TFIIF-TFIIB-TBP; +, class of partial cITC-cMed particles lacking upstream DNA-TFIIB-TBP and the mobile plank of cMed; *, class of cITC particles that do not average well with the main cITC class; †, class of Pol II-DNA/RNA particles lacking Rpb4-Rpb7. cITC-cMed(a) and cITC-cMed(b) classes from Round 3a correspond respectively to class 3 and class 4 superimposed in Fig. 5d.

The Pol II-TFIIF class (Round 6) presented with density for the TFIIF dimerization module and the Tfg1 'charged helix', but weak to no density for Tfg1 'arm' and Tfg2 'linker' regions due to the absence of upstream DNA stabilizing factors TFIIB and TBP. Classes were visualized as in **a**. **c**, 3D classification of particles from rejects of round 1 using the Pol II-DNA/RNA reconstruction as initial model. Particles were sorted into eight groups, resulting in poor 3D reconstructions. Classes were visualized as in **a**. **d**, 3D classification of particles from class 4 of round 1 using the Pol II-DNA/RNA reconstruction as initial model. Particles were sorted into four groups, resulting in EM reconstructions of cITC and cITC-cMed. These results suggest high data quality, and further the presence of a single detectable cITC-cMed conformation in the cryo-EM data, even in the absence of a cITC-cMed reference. Classes were visualized as in **a**.



Extended Data Figure 9 | Negative stain and cryo-EM reconstructions of Pol II-DNA/RNA, cITC, and cITC-cMed complexes. **a**, Four views of the negative-stain tomography reconstruction of the cITC-cMed related by 90° rotation, starting from the previously defined front view of Pol II⁴⁷. **b**, Comparison of five reference-free 2D class averages calculated from all particles used in the final negative-stain single-particle reconstruction with corresponding forward projections of the reconstruction. **c**, Orientational distribution plot of all particles in the final negative-stain single-particle reconstruction. The estimated angular accuracy is 3.2°. **d**, Fourier shell correlation of the final negative-stain single-particle cITC-cMed reconstruction (0.143 FSC = 25.2 Å resolution bin). **e**, Four views of the negative-stain single-particle reconstruction of the cITC-cMed related by 90° rotation, starting from the previously defined front view of Pol II⁴⁷.

f, Comparison of five reference-free 2D class averages calculated from all particles used in the final Pol II-DNA/RNA cryo-EM single-particle reconstruction with corresponding forward projections of the reconstruction. **g**, Orientational distribution plot of all particles in the final cryo-EM Pol II-DNA/RNA single-particle reconstruction. The estimated angular accuracy is 3.2°. **h**, Fourier shell correlation of the final Pol II-DNA/RNA cryo-EM single-particle reconstruction (FSC = 0.143). **i**, Two views of the Pol II-DNA/RNA cryo-EM map are shown from the previously defined front view of Pol II⁴⁷ and rotated by 180°, and are coloured by local resolution. **j**, Two views of the Pol II-DNA/RNA cryo-EM map are shown from the previously defined front view of Pol II⁴⁷ and rotated by 180°, and are coloured by variance (the standard deviation, StdDev, of the normalized intensity value). **k-o**, As f-j but for the cITC reconstruction. **p-t**, As f-j but for the cITC-cMed reconstruction.

Extended Data Table 1 | Components of the cITC–cMed complex

	Protein subunits	Length (aa)	Molecular weight (kDa)
RNA polymerase II	Rpb1	1733	191.6
	Rpb2	1224	138.8
	Rpb3 [‡]	318	35.3
	Rpb4	221	25.4
	Rpb5	215	25.1
	Rpb6	155	17.9
	Rpb7	171	19.1
	Rpb8	146	16.5
	Rpb9	122	14.3
	Rpb10	70	8.3
	Rpb11	120	13.6
	Rpb12	70	7.7
General initiation factors	TFIIB [‡]	345	38.2
	TBP [‡]	61–240	20.2
	Tfg1* [‡]	734	82.3
	Tfg2	400	46.6
Mediator head module	Med6	295	32.8
	Med8	222	25.3
	Med11	115	13.3
	Med17	687	78.5
	Med18	307	34.3
	Med20	210	22.9
	Med22	121	13.7
Mediator middle module	Med4	284	32.2
	Med7	222	25.6
	Med9	149	17.4
	Med10	157	17.9
	Med21	140	16.1
	Med31	127	14.7
Mediator (other subunits)	Med14 [†]	745	84.6
	Med19	220	24.9
Total	31 subunits	10,220	1,155.1
Nucleic acid strands		Length (nt)	Molecular weight (kDa)
Template	DNA	72	22.2
Non-template	DNA	72	22.5
Initial transcript	RNA	6	2.1

aa, amino acids; nt, nucleotides; kDa, kilodalton.

*Tfg1 was from *S. mikatae*⁹⁷.

[†]Med14 construct contains an additional N-terminal 10×His tag.

[‡]Constructs contain an N- or C-terminal 6×His tag as described⁹.

Explosive lithium production in the classical nova V339 Del (Nova Delphini 2013)

Akito Tajitsu¹, Kozo Sadakane², Hiroyuki Naito^{3,4}, Akira Arai^{5,6} & Wako Aoki⁷

The origin of lithium (Li) and its production process have long been uncertain. Li could be produced by Big Bang nucleosynthesis, interactions of energetic cosmic rays with interstellar matter, evolved low-mass stars, novae, and supernova explosions. Chemical evolution models and observed stellar Li abundances suggest that at least half the Li may have been produced in red giants, asymptotic giant branch (AGB) stars, and novae^{1–3}. No direct evidence, however, for the supply of Li from evolved stellar objects to the Galactic medium has hitherto been found. Here we report the detection of highly blue-shifted resonance lines of the singly ionized radioactive isotope of beryllium, ⁷Be, in the near-ultraviolet spectra of the classical nova V339 Del (Nova Delphini 2013) 38 to 48 days after the explosion. ⁷Be decays to form ⁷Li within a short time (half-life of 53.22 days⁴). The ⁷Be was created during the nova explosion via the alpha-capture reaction ³He(α , γ)⁷Be (ref. 5). This result supports the theoretical prediction that a significant amount of ⁷Li is produced in classical nova explosions.

V339 Del (Nova Delphini 2013) is a classical nova that was discovered as a bright 6.8-magnitude (unfiltered) source on 2013 August 14.584 Universal Time (UT)⁶. 40 h after the discovery, a maximum was reached on August 16.25 (Modified Julian Day (MJD) = 56,520.25) at 4.3 magnitude in the Johnson V-band⁷. Then, it began a normal decline.

High-resolution spectra (resolving power $R = 90,000$ – $60,000$) of V339 Del were obtained at four epochs after its outburst (+38 d, +47 d, +48 d, and +52 d). These spectra contain a series of broad emission lines originating from neutral hydrogen (H I, Balmer series) and other permitted transitions of neutral or singly ionized species (for example, Fe II, He I and Ca II). These emission lines are usually seen in post-outburst spectra of classical novae. Most of these broad emission lines are accompanied by sharp and blue-shifted multiple absorption lines at their blue edges. The typical radial velocity (v_{rad}) of these highly blue-shifted absorption lines is about $-1,000 \text{ km s}^{-1}$. Figure 1a and b display the spectrum obtained at +47 d in the vicinities of the H η and Ca II K lines, respectively. The H I line is accompanied by a broad emission with a full-width at half-maximum (FWHM) of $\sim 1,300 \text{ km s}^{-1}$ centred at a radial velocity of $v_{\text{rad}} \approx 0 \text{ km s}^{-1}$. The Ca II K (and H) has a weak but broad emission and strong absorption components caused by the interstellar absorptions. In addition to these profiles around the rest positions of both lines, two sharp absorption components are found at $v_{\text{rad}} = -1,268 \pm 2 \text{ km s}^{-1}$ and $-1,103 \pm 1 \text{ km s}^{-1}$. The latter seems to be stronger than the former. Such absorption line systems have been found in post-outburst spectra of several classical novae^{8,9}. The absorption line systems in V339 Del contain numerous transitions originating from singly ionized iron-group species (Fe II, Ti II, Cr II, Mn II and Ni II). The depths of all blue-shifted absorption lines in V339 Del are only about 25% or less of the continuum level, while the bottoms of some strong lines (for example, the Balmer lines; see Extended Data Fig. 2) have flat features, suggesting that the absorption is saturated. These observational results can be interpreted as the effect of absorbing materials partially covering the background light source. There are no Na I D doublet lines (at

588.995 nm and 589.592 nm), which are often found to be the strongest absorption features in novae within a few weeks after their outbursts^{8,10}. We interpret this as indicating that the ejected gas has evolved into a higher ionization stage before our observing epochs (5–7 weeks after the explosion). The observed spectral energy distribution of this nova indicates that the shape of the continuous radiation had entered a very hot stage (effective temperature $> 100,000 \text{ K}$) within 5 weeks after the explosion¹¹. Other observed characteristics of this nova (for example, light curves, optical and ultraviolet emission lines) show that it is a typical Fe II nova with a carbon–oxygen (CO) white dwarf^{12,13}.

In these absorption line systems, we noticed two remarkable pairs of absorption features near 312 nm. These correspond to the absorption components originating from transitions at $\sim 313 \text{ nm}$. These pairs are marked A, B and C, D, respectively, in Fig. 1c. Adopting the wavelengths of the resonance doublet lines^{14,15} of singly ionized ⁷Be at 313.0583 nm and 313.1228 nm, we find that features A and B coincide with the $-1,103 \text{ km s}^{-1}$ components of the 313.0583 nm and 313.1228 nm lines, respectively. Similarly, features C and D coincide with the $-1,268 \text{ km s}^{-1}$ components of these two lines. Separations in wavelength between features A and B and between features C and D are consistent with the separation between the doublet lines within the measurement uncertainties. Figure 1d illustrates these coincidences on the velocity scale. The high resolution of the spectrum ($\sim 0.0052 \text{ nm}$) means that we can clearly distinguish them from the doublet of ⁹Be II at 313.0422 nm and 313.1067 nm (ref. 14). After ruling out the possibility of alternative identifications, we conclude that these absorption features at 312 nm are caused by ⁷Be, and not by ⁹Be (the only stable isotope of Be). Original ⁹Be contained in the progenitor star would have been depleted during its evolution because this isotope is destroyed at temperatures $T > 3 \times 10^6 \text{ K}$. On the other hand, production of the unstable isotope ⁷Be by the reaction ³He(α , γ)⁷Be in nova explosions has been theoretically predicted^{16–21}.

The transition probability of the ⁷Be II line at 313.0583 nm ($\log(gf) = -0.178$, where g is the statistical weight of the lower level and f is the oscillator strength of the transition) is twice as large as that of the ⁷Be II line at 313.1228 nm ($\log(gf) = -0.479$)¹⁴. Owing to saturation effects, the ratio of their equivalent widths is expected to be in the range between 2 (no saturation) and 1 (complete saturation). The measured ratios are 1.1 ± 0.3 and 1.6 ± 0.4 for the components at $v_{\text{rad}} = -1,268 \text{ km s}^{-1}$ and $-1,103 \text{ km s}^{-1}$, respectively. These are within the range expected for the doublet, although the values contain some errors (within $\pm 25\%$), mainly because of the uncertainty in the continuum placement. The weaker component at $v_{\text{rad}} = -1,268 \text{ km s}^{-1}$ has a ratio closer to complete saturation. This can be interpreted as resulting from the fact that the absorbing gas cloud moving with $v_{\text{rad}} = -1,268 \text{ km s}^{-1}$ has a smaller covering factor, and at the same time, a higher column density of ⁷Be ions than the gas cloud moving with $v_{\text{rad}} = -1,103 \text{ km s}^{-1}$.

Figure 2 displays the velocity plots of normalized spectra for different species of the absorption line systems at four observing epochs from +38 d to +52 d after the explosion. On day +38, the ⁷Be II doublet has

¹Subaru Telescope, National Astronomical Observatory of Japan, 650 North A'ohoku Place, Hilo, Hawaii 96720, USA. ²Astronomical Institute, Osaka Kyoiku University, Asahigaoka, Kashiwara, Osaka 582-8582, Japan. ³Graduate School of Science, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8602, Japan. ⁴Nayoro Observatory, 157-1 Nisshin, Nayoro, Hokkaido 096-0066, Japan. ⁵Center for Astronomy, University of Hyogo, Sayo-cho, Hyogo 679-5313, Japan. ⁶Koyama Astronomical Observatory, Kyoto Sangyo University, Motoyama, Kamigamo, Kita-ku, Kyoto 603-8555, Japan. ⁷National Astronomical Observatory of Japan, 2-21-1 Osawa, Mitaka, Tokyo 181-8588, Japan.

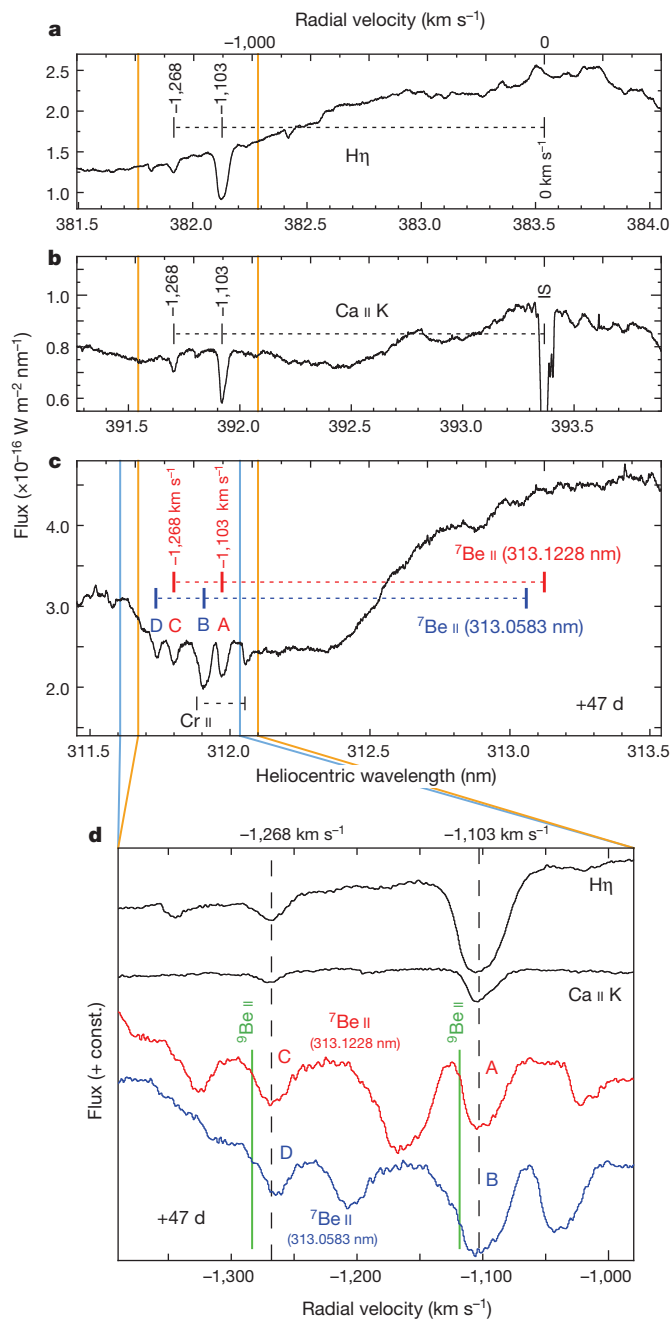


Figure 1 | Blue-shifted absorption line systems in the spectrum of V339 Del obtained at day +47. **a–c,** The spectrum in the vicinity of H γ (**a**), Ca II K (**b**), and the $^7\text{Be II}$ doublet (**c**), on the velocity (upper horizontal) scale. Two blue-shifted absorption components and the zero velocity position for each line are indicated. **b,** The strong interstellar (IS) absorption of Ca II K is centered at $v_{\text{rad}} \approx 0 \text{ km s}^{-1}$. **c,** The velocity scale is adjusted to one of the $^7\text{Be II}$ doublet (313.1228 nm, red). The positions of blue-shifted components of Cr II at 313.205 nm are displayed at the bottom. **d,** The enlarged radial velocity profiles. The vertical dashed lines show two common absorption components. The expected positions of the $^9\text{Be II}$ doublet are indicated by green lines.

an absorption component at $v_{\text{rad}} = -1,386 \pm 3 \text{ km s}^{-1}$ and shows a complicated profile near $v_{\text{rad}} \approx -1,000 \text{ km s}^{-1}$. From day +47 to day +48, the absorption components at $v_{\text{rad}} = -1,268 \text{ km s}^{-1}$ and $-1,103 \text{ km s}^{-1}$ on day +47 shift by $-26 \pm 3 \text{ km s}^{-1}$ and $-17 \pm 4 \text{ km s}^{-1}$ blueward, respectively. These changes can be interpreted as due to the fact that we are observing accelerating blobs of nova ejecta. All of the blue-shifted absorption line systems had disappeared in the spectrum of day +52 except for the meta-stable He I lines at 318.8 nm and 388.9 nm. This indicates that

the gas in the absorption line systems has evolved into a higher ionization stage, as discussed in the case of the nova V1280 Sco (ref. 22).

These observations show that: (1) several blue-shifted absorption lines with different velocities are found from different species at each epoch; (2) radial velocities of different transitions belonging to a velocity component determined by Gaussian fittings agree within $|\Delta v_{\text{rad}}| < 1\text{--}3 \text{ km s}^{-1}$; (3) each component shifts blueward with time, indicating that the ejecta is being accelerated; (4) the strengths of the blue-shifted absorption lines weaken quickly during the observing period. The velocities and the strengths of the $^7\text{Be II}$ doublet are perfectly synchronized with those of other species. This means that the gas producing the absorption line systems of V339 Del contains a considerable amount of ^7Be ion, which can produce detectable absorption lines and strongly suggests that the gas must have experienced an explosive thermonuclear runaway on the surface of the white dwarf.

Our spectroscopic detection of ^7Be in a classical nova immediately connects to the production of ^7Li . The production of ^7Be via the nuclear reaction $^3\text{He}(\alpha, \gamma)^7\text{Be}$ in novae has been studied theoretically. However, no observational confirmation has been made. This is because ^7Be is very transient and is only observable in the near-ultraviolet range, where the atmospheric absorption severely obstructs observations from ground-based telescopes. In the case of V339 Del, the ^7Be doublet can be identified only within a very short period (from ~ 6 weeks to ~ 7 weeks after the outburst). In earlier epochs, saturation effects might make it difficult to identify. For nearby bright novae, there have been several attempts to detect the 478 keV gamma-ray line produced by the decay of ^7Be . However, no definite detection has been reported because of the insufficient sensitivity^{23,24}. The ^7Be absorption lines in the near-ultraviolet spectrum of V339 Del are found in highly blue-shifted ($\sim 1,000 \text{ km s}^{-1}$) flows that have been blown off by the outburst. This means that it will soon decay to ^7Li in cooler interstellar or circumstellar matter on a time-scale given by the half-life of ^7Be (53.22 days). The absence of the $^7\text{Li I}$ line at 670.8 nm in our spectra can be interpreted as due to the fact that all of the Li in the absorbing material of V339 Del has been ionized during the observing period, as mentioned above. This is in accordance with the fact that no Na I D lines are found in the absorption line system.

The $^7\text{Be II}$ doublet corresponds to the doublet of the Ca II resonance lines (H and K lines) on the atomic energy level diagrams. The Ca II K line at 393.366 nm has $\log(gf) = +0.135$ (ref. 14). Supposing that most of the ^7Be and Ca in the absorption line system is singly ionized and the resonance lines of both ions are unsaturated, the ratio of their equivalent widths directly reflects the number density ratio between ^7Be and Ca ions. In the spectrum obtained at +47 d, the ratios of $^7\text{Be II}$ (313.1228 nm)/Ca II K, which are less affected by saturation and/or contamination, are $\sim 1.3 \pm 0.3$ and $\sim 0.7 \pm 0.2$ for the $-1,268 \text{ km s}^{-1}$ and the $-1,103 \text{ km s}^{-1}$ components, respectively. This means that the column number density of ^7Be is 5.3–2.9 times higher than that of Ca. Using this $^7\text{Be}/\text{Ca}$ number ratio, the mass fraction of ^7Be relative to the sum of all the constituent mass components, $X(^7\text{Be})$, can be presented as: $X(^7\text{Be}) = (4.4 \pm 2.2) \times 7/40 \times X(\text{Ca})$. If we adopt the solar value of $X(\text{Ca}) = 10^{-4.19}$, for instance, the $X(^7\text{Be})$ in the absorbing gas system should amount to $\sim 10^{-4.3 \pm 0.3}$. The error estimation includes the uncertainty in the local continuum placement (less than $\pm 25\%$) and the difference of equivalent widths derived from individual velocity components ($\sim \pm 30\%$). Several additional factors, such as the abundance of Ca in the absorbing gas, and the difference in the ionization state between ^7Be and Ca, are difficult to estimate because the nova ejecta model is not yet established. Taking such uncertainties into account, the error involved in the above estimate of the ^7Be abundance could be several times larger. However, in spite of the large uncertainty, the abundance of ^7Be is larger than, or at least as large as, theoretical predictions for CO novae (for example, $X(^7\text{Be}) \approx 10^{-5.1}$ or less; ref. 21). This indicates that classical novae could have an important role as contributors to Galactic ^7Li .

The observed ^7Li evolutionary curve³ shows a plateau for young Galactic ages (less than about 2.5 Gyr), followed by a steep rise. To explain this, we require a relatively low-mass stellar component that evolves

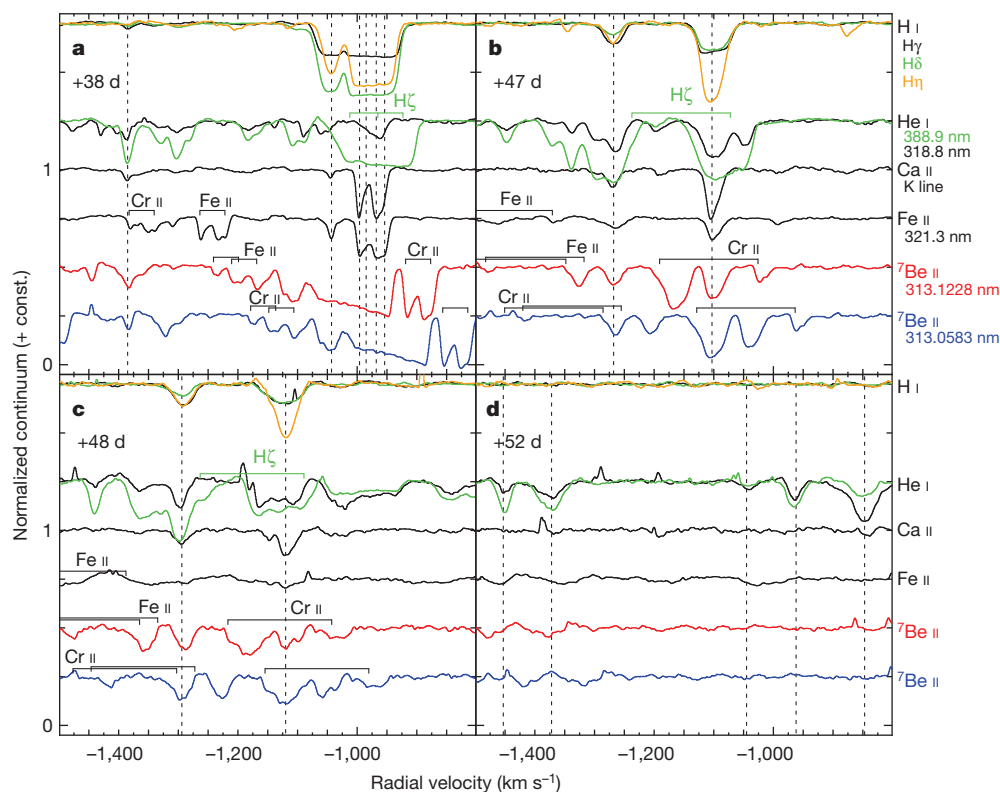


Figure 2 | Time variations of the blue-shifted absorption line systems from day +38 to day +52. Absorption line systems originating from different species at four observing epochs are plotted on the velocity scale. All lines are normalized to the local continuum. Blue-shifted absorption components observed at each epoch are indicated by vertical dashed lines. The identified line

or expected line contaminations are labelled above each lines. **a**, On day +38, the profile of the ${}^7\text{Be II}$ doublet around $v_{\text{rad}} \approx -1,000 \text{ km s}^{-1}$ is complicated, and possibly interpreted as being saturated. **d**, On day +52, no blue-shifted absorption can be found except for the metastable He I lines.

over a long lifetime. Candidates for this, such as low-mass red giants or novae, have been proposed to be major sources of ${}^7\text{Li}$ production (more than half of the Solar System Li, as measured in meteorites) in the Galaxy^{1–3}. The production of ${}^7\text{Li}$ in low-mass stars has been theoretically studied^{25–29}, and Li-enhanced red giants and AGB stars have indeed been identified³⁰. The contribution to Li enrichment in the Galaxy by these objects has, however, not been confirmed. This is because the Li-rich phase in these stars might be of quite limited duration and the contribution is dependent upon the mass-loss rate of such objects. Nova eruptions involve a long delay time before working as stellar ${}^7\text{Li}$ factories. This is because ${}^3\text{He}$ -rich low-mass secondaries are necessary to produce ${}^7\text{Be}$ efficiently via the ${}^3\text{He}(\alpha, \gamma){}^7\text{Be}$ reaction¹⁸. It is important to know whether this phenomenon is common among classical novae to quantify their contribution to the rapid increase of ${}^7\text{Li}$ in the Galaxy. Since V339 Del appears to be one of the ordinary Fe II type novae that occupy $\sim 60\%$ of all classical novae¹², the ${}^7\text{Be}$ production found in this object might be occurring in many classical novae. Our successful detection of ${}^7\text{Be}$ in V339 Del indicates that measurements of the ${}^7\text{Be}$ lines in the near-ultraviolet range for post-outburst novae within the lifetime of this isotope is a powerful way to estimate the contribution of novae to the chemical evolution of lithium in the Galaxy.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 2 September; accepted 16 December 2014.

- Romano, D., Matteucci, F., Molaro, P. & Bonifacio, P. The galactic lithium evolution revisited. *Astron. Astrophys.* **352**, 117–128 (1999).
- Romano, D., Matteucci, F., Ventura, P. & D'Antona, F. The stellar origin of ${}^7\text{Li}$. Do AGB stars contribute a substantial fraction of the local Galactic lithium abundance? *Astron. Astrophys.* **374**, 646–655 (2001).
- Prantzos, N. Production and evolution of Li, Be, and B isotopes in the Galaxy. *Astron. Astrophys.* **542**, A67 (2012).

- Audi, G., Bersillon, O., Blachot, J. & Wapstra, A. H. The NUBASE evaluation of nuclear and decay properties. *Nucl. Phys. A* **729**, 3–128 (2003).
- Cameron, A. G. W. & Fowler, W. A. Lithium and the s-PROCESS in Red-Giant Stars. *Astrophys. J.* **164**, 111–114 (1971).
- Waagen, E. O. Nova Delphini 2013 = PNV J20233073+2046041. *AAVSO Alert Notice* **489** (2013).
- Munari, U. *et al.* After a post-maximum plateau Nova Del 2013 has begun a normal decline. *Astron. Telegr.* **5304**, 1 (2013).
- Williams, R., Mason, E., Della Valle, M. & Ederoclite, A. Transient heavy element absorption systems in novae: episodic mass ejection from the secondary star. *Astrophys. J.* **685**, 451–462 (2008).
- Sadakane, K., Tajitsu, A., Mizoguchi, S., Arai, A. & Naito, H. Discovery of multiple high-velocity narrow circumstellar Na I lines in Nova V1280 Sco. *Publ. Astron. Soc. Jpn* **62**, L5–L10 (2010).
- McLaughlin, D. B. in *Stellar Atmospheres* (ed. Greenstein, J. L.) 585–652 (The University of Chicago Press, 1960).
- Skopal, A. *et al.* Early evolution of the extraordinary Nova Delphini 2013 (V339 Del). *Astron. Astrophys.* **569**, A112 (2014).
- Williams, R. E. The formation of novae spectra. *Astron. J.* **104**, 725–733 (1992).
- Warner, B. Cataclysmic variable stars. *Camb. Astrophys. Ser.* **28**, 257–306 (1995).
- Kramida, A., Ralchenko, Yu., Reader, J. & the NIST ASD team. *NIST Atomic Spectra Database Ver. 5.1* <http://physics.nist.gov/asd> (National Institute of Standards and Technology, 2013).
- Yan, Z.-C., Nörtershäuser, W. & Drake, G. W. F. High precision atomic theory for Li and Be⁺: QED shifts and isotope shifts. *Phys. Rev. Lett.* **100**, 243002 (2008).
- Arnould, M. & Norgaard, H. The explosive thermonuclear formation of ${}^7\text{Li}$ and ${}^{11}\text{B}$. *Astron. Astrophys.* **42**, 55–70 (1975).
- Starrfield, S., Truran, J. W., Sparks, W. M. & Arnould, M. On Li-7 production in nova explosions. *Astrophys. J.* **222**, 600–603 (1978).
- D'Antona, F. & Matteucci, F. Galactic evolution of lithium. *Astron. Astrophys.* **248**, 62–71 (1991).
- Boffin, H. M. J., Paulus, G., Arnould, M. & Mowlavi, N. The explosive thermonuclear formation of Li-7 revisited. *Astron. Astrophys.* **279**, 173–178 (1993).
- Hernanz, M., Jose, J., Coc, A. & Isern, J. On the synthesis of ${}^7\text{Li}$ and ${}^7\text{Be}$ in novae. *Astrophys. J.* **465**, L27–L30 (1996).
- José, J. & Hernanz, M. Nucleosynthesis in classical novae: CO versus ONe white dwarfs. *Astrophys. J.* **494**, 680–690 (1998).
- Naito, H., Tajitsu, A., Arai, A. & Sadakane, K. Discovery of metastable helium absorption lines in V1280 Scorpii. *Publ. Astron. Soc. Jpn* **65**, 37 (2013).

23. Harris, M. J. *et al.* Transient gamma-ray spectrometer observations of gamma-ray lines from novae. III. The 478 keV line from ${}^7\text{Be}$ decay. *Astrophys. J.* **563**, 950–957 (2001).
24. Hernanz, M. in *Classical Novae* (eds Bode, M. F. & Evans, A.) 2nd edn, 252–284 (Cambridge Astrophys. Ser. 43, Cambridge University Press, 2008).
25. Sackmann, I.-J. & Boothroyd, A. I. Creation of ${}^7\text{Li}$ and destruction of ${}^3\text{He}$, ${}^9\text{Be}$, ${}^{10}\text{B}$, and ${}^{11}\text{B}$ in low-mass red giants, due to deep circulation. *Astrophys. J.* **510**, 217–231 (1999).
26. de la Reza, R., da Silva, L., Drake, N. A. & Terra, M. A. On ${}^7\text{Li}$ enrichment by low-mass metal-poor red giant branch stars. *Astrophys. J.* **535**, L115–L117 (2000).
27. Sackmann, I.-J. & Boothroyd, A. I. The creation of super-rich lithium giants. *Astrophys. J.* **392**, L71–L74 (1992).
28. Travaglio, C. *et al.* Galactic chemical evolution of lithium: interplay between stellar sources. *Astrophys. J.* **559**, 909–924 (2001).
29. Ventura, P. & D'Antona, F. The role of lithium production in massive AGB and super-AGB stars for the understanding of multiple populations in globular clusters. *Mon. Not. R. Astron. Soc.* **402**, L72–L76 (2010).
30. Melo, C. H. F. *et al.* On the nature of lithium-rich giant stars. Constraints from beryllium abundances. *Astron. Astrophys.* **439**, 227–235 (2005).

Acknowledgements This work is based on data collected at the Subaru Telescope, which is operated by the National Astronomical Observatory of Japan (NAOJ). We acknowledge with thanks the variable star observations from the AAVSO International Database contributed by observers worldwide and used in this research.

Author Contributions A.T. planned and carried out the Subaru High Dispersion Survey observations, reduced and analysed the data and prepared the manuscript. K.S., H.N., A.A., and W.A. participated in the discussion and contributed to the process of manuscript preparation.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.T. (tajitsu@naoj.org).

METHODS

Discovery. V339 Del (Nova Delphini 2013) is a classical nova that was discovered as a bright 6.8 magnitude (unfiltered) source by Koichi Itagaki on 2013 August 14.584 UT and announced in the American Association of Variable Star Observers (AAVSO) Alert Notice⁶. Its progenitor is estimated to be a blue star (USNO B-1 1107-0509795) with 17.20 magnitude in the Johnson B-band, and 17.45 magnitude in the Cousins R_C -band on the first Palomar Sky Survey Plates (exposed on 1951 July 7), and with $B \approx 17.39$, $R_C \approx 17.74$ on the second Palomar Sky Survey plates (exposed on 1990 July 18 and September 15, respectively)³¹. No significant changes were found in its photometric behaviour for at least a few years before the outburst³². On an unfiltered pre-discovery image obtained on 2013 August 13.998 UT, the object was still at 17.1 mag (ref. 33). This means that the object was still in quiescence until at least 14 h before its discovery, and that it showed a very fast rise to the maximum. After 40 h from the discovery, maximum was reached on Aug 16.25 (MJD = 56,520.25) at $V = 4.3$ (ref. 7). Then, it began a normal decline. The nova had been detected as a transient high-energy gamma-ray (>100 MeV) source within about 10 days after the outburst³⁴. Angular sizes of the expanding shell around the nova had been monitored until about day +40 using near-infrared interferometric observations³⁵. Then, incorporating the expansion velocity obtained in the optical region, the distance to the nova was derived as 4.54 ± 0.59 kiloparsecs from the Sun.

Observations and data reduction. The post-outburst spectra of V339 Del were obtained using the High Dispersion Spectrograph (HDS)³⁶ of the 8.2 m Subaru Telescope at four epochs from 2013 September 23 to October 7 (+38 d, +47 d, +48 d and +52 d after the maximum). According to the AAVSO light curves (see Extended Data Fig. 1), our first observation was just before the start of the rapid decline in optical magnitudes by dust formation³⁷. The following three were obtained during the continuous decline. We obtained spectra under three configurations of the spectrograph, which cover the wavelength regions from 303 nm to 463 nm, from 411 nm to 686 nm, and from 667 nm to 936 nm. The spectral resolving power was set to $R \approx 90,000$ or 60,000 with 0.4" (0.2 mm) or 0.6" (0.3 mm) slit widths, respectively. The exact times and wavelength ranges of obtained spectra are summarized in Extended Data Table 1. Data reduction was carried out using the Image Reduction and Analysis Facility (IRAF) software in a standard manner. The nonlinearity in the detectors was corrected by the method described in ref. 38. The wavelength calibration was performed using a Th-Ar comparison spectrum and the typical residual in wavelength calibration is about 10^{-4} nm (~ 0.1 km s⁻¹) or less for each spectrograph configuration. The typical systemic variance of the spectrograph is about 10^{-4} nm or less per hour. We also examined the accuracy of radial velocity determination in our measurement using the identified iron-group transitions in the range 315–351 nm. For the spectrum obtained at day +38, the velocity of the strongest component in the absorption line system was -996.1 ± 0.7 km s⁻¹, determined by Gaussian fittings. In total, we concluded that the residual in our velocity scale determination was ± 1 km s⁻¹. Spectrophotometric calibration was performed using the spectrum of BD +28°4211 (ref. 39), obtained at nearly the same nova altitude on the same nights. All spectra were converted to the heliocentric scale. A correction for interstellar extinction has not been applied. The average signal-to-noise ratio in the spectra obtained at four epochs is ~ 140 at ~ 312 nm, where we found the ⁷Be lines.

Highly blue-shifted absorption line system. The spectra of V339 Del exhibit a series of broad Fe II emission lines, which indicate that the object is a typical Fe II type nova¹². Since no strong emission originating from Ne is found in the spectrum even at day +52, the white dwarf in the system is supposed to be a CO white dwarf³.

Extended Data Fig. 2a displays the radial velocity profiles of three Fe II belonging to the same multiplet number⁴⁰ (42) in the spectrum of day +38. The absorption line system on day +38 clearly consists of five components at $v_{\text{rad}} = -954$ km s⁻¹, -968 km s⁻¹, -985 km s⁻¹, -996 km s⁻¹, and $-1,043$ km s⁻¹, as indicated by the dashed lines in Extended Data Fig. 2b. Similarly, blue-shifted absorption components are found in Balmer lines (Extended Data Fig. 2c) and also in other permitted lines (Ca II H and K, He I at 587.6 nm). In the near-ultraviolet range, numerous absorption lines in the complex continuum are identified as the transitions of singly ionized iron-group species. Most of them belong to the absorption line systems found in the visual region (Extended Data Fig. 3). We applied a Doppler correction using the radial velocity of the strongest blue-shifted absorption line in the system to identify sources of transitions (see Extended Data Fig. 3b). We use the velocities for Doppler corrections as $v_{+38} = -996$ km s⁻¹, $v_{+47} = -1,103$ km s⁻¹, and $v_{+48} = -1,120$ km s⁻¹ for days +38, +47, and +48, respectively (in Extended Data Figs 3b and 4a–c). All of the identified transitions originate from levels of low excitation potential (<4 eV). The residual intensity at the bottom of these lines exceeds 75% of the continuum, whereas the bottoms of some strong Fe II and Balmer lines show flat structures. These observations suggest that the saturated absorption lines are created by clouds of absorbing gas, which cover only about 25% of the continuum-emitting region.

Very similar short-lived blue-shifted metallic absorption systems have been reported in post-outburst spectra of several classical novae (such as the Transient

Heavy Element Absorption; THEA⁸). In particular, the great majority of novae show strong blue-shifted (400 – $1,000$ km s⁻¹) multiple absorption components in the Na I D doublet in the days following their outbursts. In the very slow nova V1280 Sco, multiple high-velocity (700 – 900 km s⁻¹) absorption components were found in the Na I D doublet even 800 days after its maximum⁹. Although no absorption component of the Na I D is found at any epoch of our observations of V339 Del, other characteristics of the absorption line systems found in V339 Del are quite similar to those of the THEAs in other novae. They are (1) highly blue-shifted ($\sim 1,000$ km s⁻¹), (2) divided into several velocity components, (3) time variable in their shapes and velocities, and (4) short-lived (2–8 weeks).

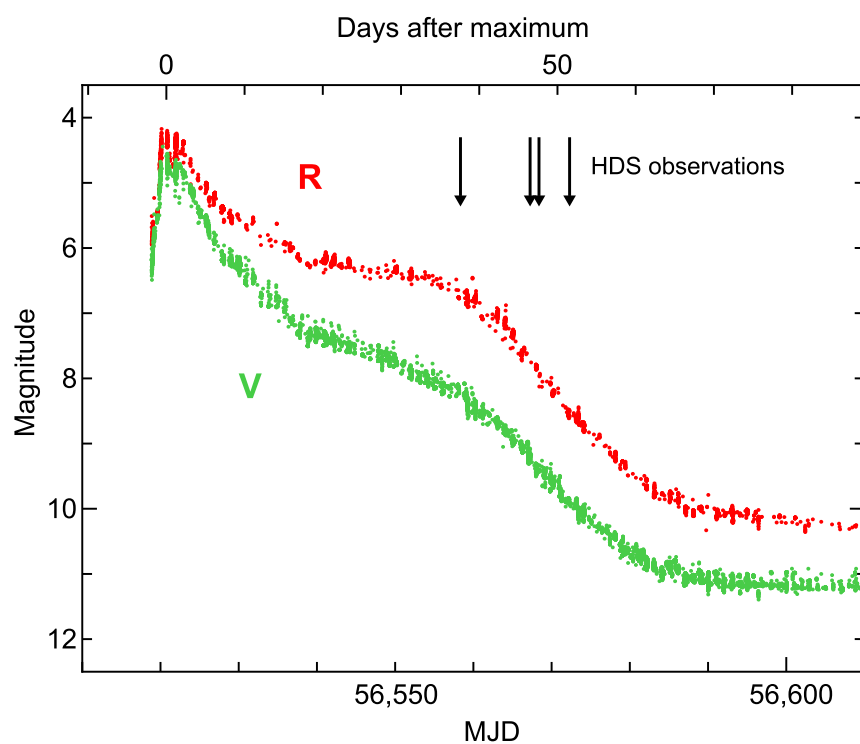
Contamination of the ⁷Be II doublet. We carefully inspected possible contaminations of absorption lines originating from other species to the ⁷Be II lines, consulting the atomic line database⁴¹. Extended Data Fig. 4 displays the spectra in the vicinity of the ⁷Be II doublet obtained at four epochs of our observations. On the spectrum obtained at day +47, there are no candidates that could contaminate the $-1,103$ km s⁻¹ or the $-1,268$ km s⁻¹ components of ⁷Be II at 313.1228 nm, which we use in our ⁷Be abundance estimation (see in Extended Data Fig. 4–b).

At this epoch, the other line of the ⁷Be doublet at 313.0583 nm may be contaminated by some lines originating from iron-group species. We estimate that the $-1,268$ km s⁻¹ component of Cr II (5) at 313.205 nm ($\log(gf) = +0.079$) may contaminate the $-1,103$ km s⁻¹ component of this ⁷Be II line. The influence of this contamination can be evaluated by applying the line strength ratio between the pair of velocity components of Cr II (5) at 312.497 nm ($\log(gf) = +0.018$) to that of Cr II (5) at 313.205 nm. It is quite small compared with the strength of the ⁷Be II line ($<5\%$). Concerning the $-1,268$ km s⁻¹ component of this ⁷Be II line, we conclude that the weak lines of Fe II (96) at 312.901 nm ($\log(gf) = -2.70$) and Cr II (5) at 312.869 nm ($\log(gf) = -0.32$) do not contaminate severely. This is because similar weak lines of Fe II (82) at 313.536 nm ($\log(gf) = -1.13$) and Cr II (5) at 313.668 nm ($\log(gf) = -0.25$) had completely disappeared until day +47. We can neglect the contamination from the V II (1) line at 313.0257 nm ($\log(gf) = -0.29$), because the other V II (1) line at 312.621 nm ($\log(gf) = -0.27$) is not detected on our spectra.

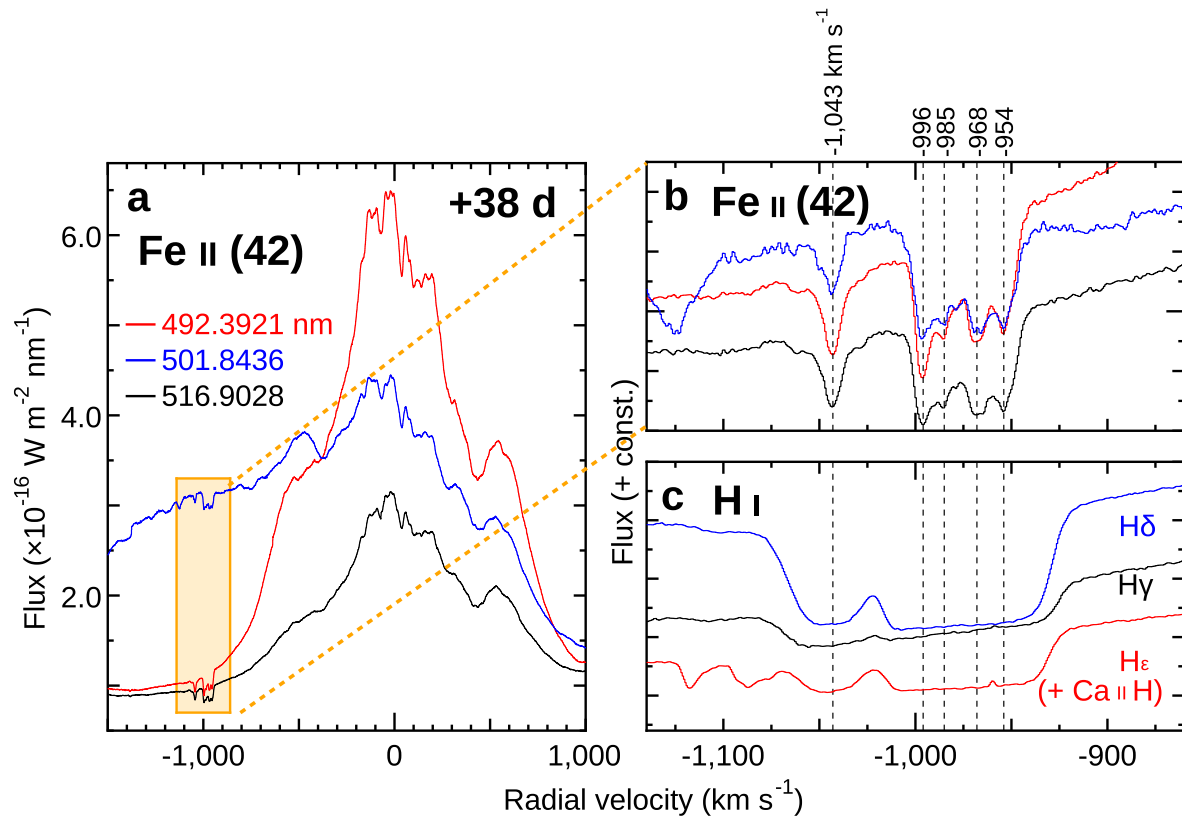
⁷Be abundance estimation. We empirically estimate the abundance of ⁷Be in the absorbing gas by comparing the equivalent widths of the ⁷Be II line with those of the Ca II K line that are the similar transitions on the atomic energy level diagrams. We assume that the covering factor of the absorbing gas cloud to the background illuminating source has no wavelength dependence. This method could be a simple and robust way to estimate the abundance ratio independently of ejecta models for nova explosions. The estimate, however, includes some uncertainties. One is the difference in the ionization potentials between Be (the first and second ionization potentials; $I_1 = 9.32$ eV, $I_2 = 18.21$ eV) and Ca ($I_1 = 6.11$ eV, $I_2 = 11.87$ eV)⁴⁴ that could result in a difference of ionization states between Be and Ca. However, all of the iron-peak elements (Ti to Fe) found in the absorption line systems that have ionization potentials ($I_1 = 6.75$ – 7.90 eV, $I_2 = 13.58$ – 18.12 eV) intermediate between those of Be and Ca, are observed only in singly ionized states, suggesting that dominant fractions of Be and Ca are in the singly ionized states, too. In the spectra obtained, we could not find any resonance lines of Sr II or Ba II, which correspond to those of Be II and Ca II. The Sr/Ca and the Ba/Ca number ratios would be quite small, as seen in the solar abundance (much less than 0.001). Another uncertainty is the $X(\text{Ca})$ in the absorption line system. Our assumption that the absorbing gas has the solar $X(\text{Ca})$ should not be far from reality, because the theoretical analysis predicts no overabundance of elements with the mass number >30 in ejecta of CO novae²¹. We remark that our ⁷Be abundance estimation is carried out using the data obtained at day +47, which is close to the half-life of ⁷Be (53.22 days). Therefore, the abundance of the freshly produced ⁷Li in this nova explosion could be twice as high as the $X(^7\text{Be})$ on day +47.

31. Munari, U. & Henden, A. Photometry of the progenitor of Nova Del 2013 (V339 Del) and calibration of a deep BVRI photometric comparison sequence. *Inform. Bull. Variable Stars* **6087**, 1 (2013).
32. Deacon, N. R. et al. Pre-outburst observations of Nova Del 2013 from Pan-STARRS 1. *Astron. Astrophys.* **563**, A129 (2014).
33. Denisenko, D. et al. V339 Delphini = Nova Delphini 2013 = Pnv J20233073+2046041. *IAU Circ. No.* **9258**, 2 (2013).
34. The Fermi-LAT Collaboration. Fermi establishes classical novae as a distinct class of gamma-ray sources. *Science* **345**, 554–558 (2014).
35. Schaefer, G. H. et al. The expanding fireball of Nova Delphini 2013. *Nature* **515**, 234–236 (2014).
36. Noguchi, K. et al. High Dispersion Spectrograph (HDS) for the Subaru Telescope. *Publ. Astron. Soc. Jpn* **54**, 855–864 (2002).
37. Shenavrin, V. I., Taranova, O. G. & Tatarnikov, A. M. Dust formation in Nova Del 2013. *Astron. Telegr.* **5431**, 1 (2013).
38. Tajitsu, A., Aoki, W., Kawanomoto, S. & Narita, N. Nonlinearity in the detector used in the Subaru Telescope High Dispersion Spectrograph. *Publ. Natl. Astron. Obs. Jpn* **13**, 1–8 (2010).
39. Massey, P., Strobel, K., Barnes, J. V. & Anderson, E. Spectrophotometric standards. *Astrophys. J.* **328**, 315–333 (1988).

40. Moore, C. E. *A Multiplet Table of Astrophysical Interest: NBS Technical Note No. 36, Reprinted Version of the 1945 edition* (US Department of Commerce, 1959).
41. Kurucz, R. & Bell, B. *Atomic Line Data* CD-ROM No. 23 (Smithsonian Astrophysical Observatory, 1995).

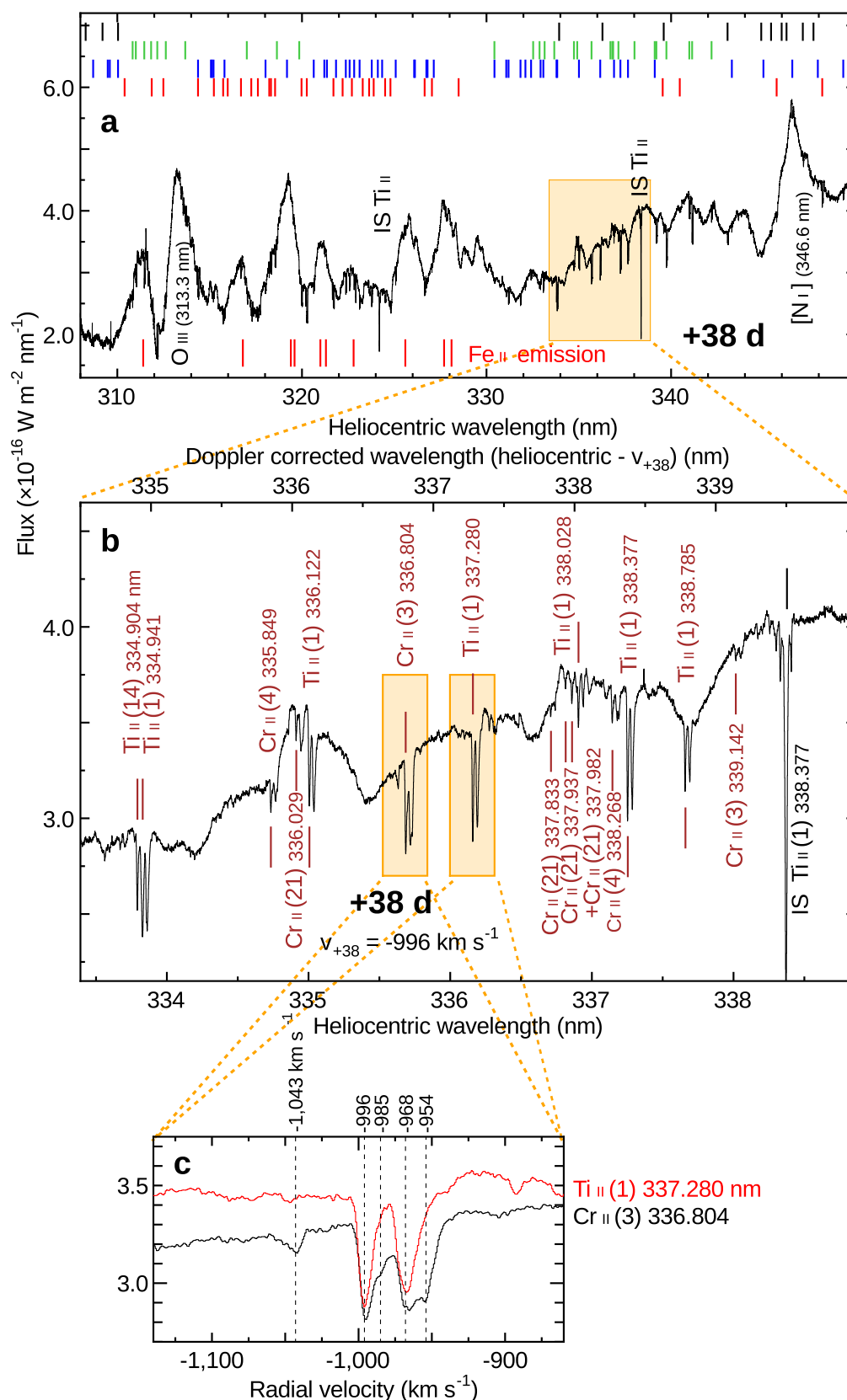


Extended Data Figure 1 | Optical light curves of V339 Del. V (green) and R (red) magnitudes are taken from the AAVSO database. The epochs of our HDS observations are indicated by arrows.



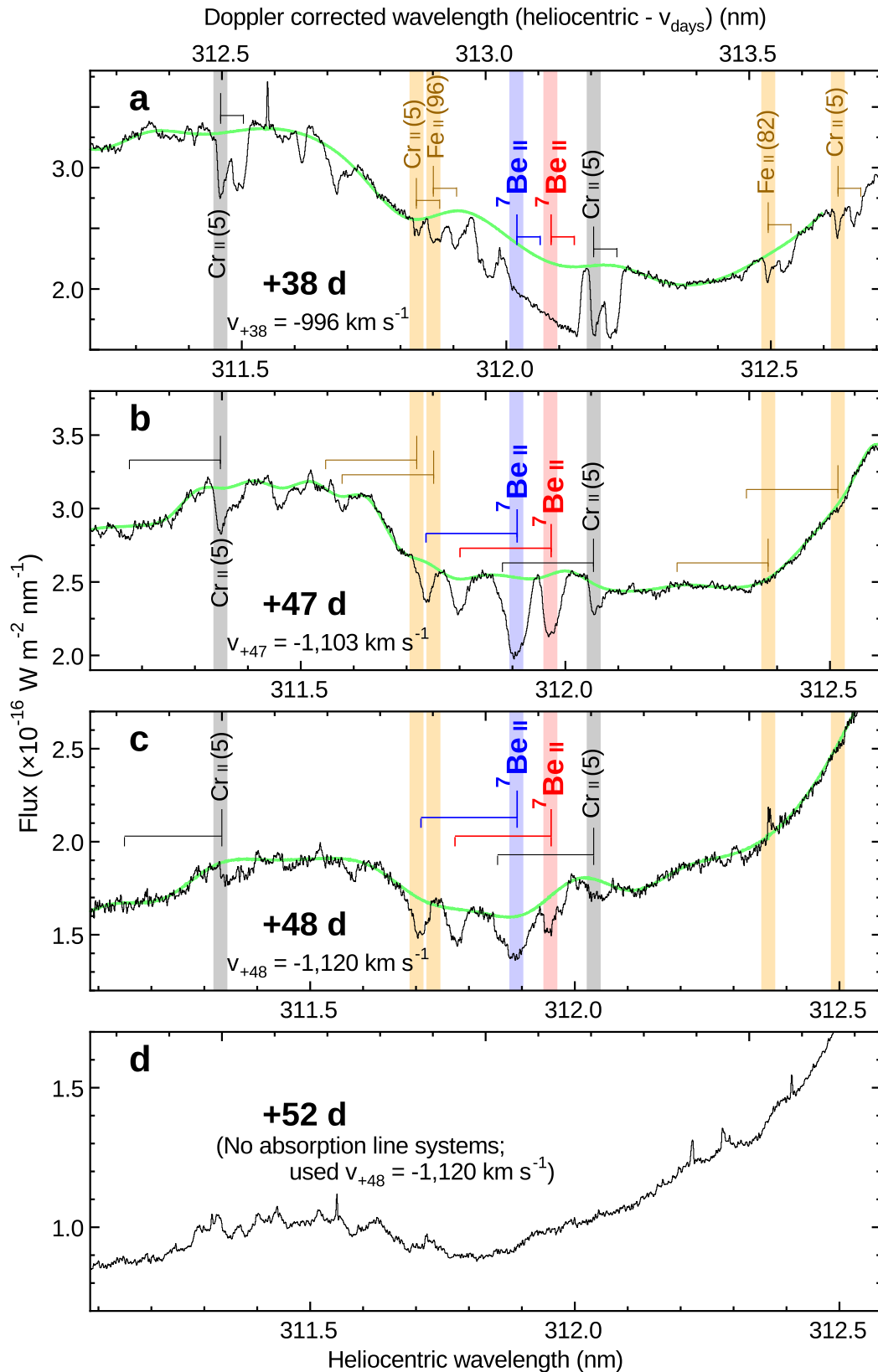
Extended Data Figure 2 | Optical spectrum of V339 Del obtained at +38 d. **a**, The radial velocity plots of three Fe II emission lines belonging to the same multiplet number³⁰ (42). In addition to the similarity of their broad emission profiles, all lines have common blue-shifted absorption line features around

$v_{\text{rad}} \approx -1,000 \text{ km s}^{-1}$. **b**, An enlarged view of the absorption line features in **a**. Dips of individual absorption line are indicated with dashed lines. **c**, The absorption line systems in H I Balmer lines drawn on the same velocity scale as in **b**.



Extended Data Figure 3 | Near-ultraviolet spectrum of V339 Del obtained at day +38. **a**, The overall view of the spectrum from 308 nm to 350 nm. Identified Fe II emission lines are indicated with red ticks at the bottom. The identified absorption line systems originating from iron-group ions are indicated by coloured ticks at the top: Fe II (red), Ti II (blue), Cr II (green), Mn II,

Ni II, and V II (black). **b**, A sample of the absorption line identification. The results of our identification are displayed along the spectrum. **c**, As for Extended Data Fig. 2b, but for two lines (Ti II and Cr II), highlighted in **b**, which are plotted on the velocity scale.



Extended Data Figure 4 | Spectra in the vicinity of the Be II doublet from day +38 to day +52. a–c, The horizontal scale is displayed with the heliocentric (bottom) and the Doppler-corrected wavelengths (top). The Doppler corrections are applied using $v_{\text{days}} = v_{+38}$, v_{+47} , and v_{+48} for panels a, b, and c, respectively. The local continua, fitted with high-order (10–20) spline

functions, are overplotted with green lines. The positions of the strongest ($v_{\text{rad}} = v_{\text{days}}$) and the second-strongest components of the absorption line system are indicated by coloured long and short lines connected by horizontal bars. d, Since no apparent absorption lines are found in day +52, a Doppler correction using v_{+48} is applied to the spectrum.

Extended Data Table 1 | Journal of HDS observations of V339 Del

Date 2013	UT* (h m)	MJD		Exposure (s)	Range (nm)	Resolution
Sep 23	6 16	56,558.261	(+38 d) [†]	720	411-686	90,000
	8 12	56,558.342		900	667-936	90,000
	10 07	56,558.423		3,000	303-463	90,000
Oct 02	5 02	56,567.210	(+47 d) [†]	3,000	303-463	60,000
	6 29	56,567.271		600	411-686	90,000
	7 18	56,567.305		900	667-936	90,000
Oct 03	9 21	56,568.390	(+48 d) [†]	3,000	303-463	60,000
Oct 07	5 05	56,572.212	(+52 d) [†]	4,800	303-463	60,000
	7 47	56,572.324		960	411-686	90,000
	8 17	56,572.346		1,500	667-936	90,000

* UT is the universal time at the start of an exposure.

†Days after the optical (*V*) maximum (MJD = 56,520.25).

Direct observation of bond formation in solution with femtosecond X-ray scattering

Kyung Hwan Kim^{1,2*}, Jong Goo Kim^{1,2*}, Shunsuke Nozawa^{3*}, Tokushi Sato^{3†*}, Key Young Oang^{1,2}, Tae Wu Kim^{1,2}, Hosung Ki^{1,2}, Junbeom Jo^{1,2}, Sungjun Park^{1,2}, Changyong Song⁴, Takahiro Sato^{4†}, Kanade Ogawa^{4†}, Tadashi Togashi⁵, Kensuke Tono⁵, Makina Yabashi⁴, Tetsuya Ishikawa⁴, Joonghan Kim⁶, Ryong Ryoo^{1,2}, Jeongho Kim⁷, Hyotcherl Ihee^{1,2} & Shin-ichi Adachi^{3,8}

The making and breaking of atomic bonds are essential processes in chemical reactions. Although the ultrafast dynamics of bond breaking have been studied intensively using time-resolved techniques^{1–3}, it is very difficult to study the structural dynamics of bond making, mainly because of its bimolecular nature. It is especially difficult to initiate and follow diffusion-limited bond formation in solution with ultrahigh time resolution. Here we use femtosecond time-resolved X-ray solution scattering to visualize the formation of a gold trimer complex, $[\text{Au}(\text{CN})_2^-]_3$ in real time without the limitation imposed by slow diffusion. This photoexcited gold trimer, which has weakly bound gold atoms in the ground state^{4–6}, undergoes a sequence of structural changes, and our experiments probe the dynamics of individual reaction steps, including covalent bond formation, the bent-to-linear transition, bond contraction and tetramer formation with a time resolution of ~ 500 femtoseconds. We also determined the three-dimensional structures of reaction intermediates with sub-ångström spatial resolution. This work demonstrates that it is possible to track in detail and in real time the structural changes that occur during a chemical reaction in solution using X-ray free-electron lasers⁷ and advanced analysis of time-resolved solution scattering data.

The functional efficiencies of photoactive molecules are governed by long-lived electronic excited states that are directly involved in functional transitions of the molecules. In fact, early stages of their photoinduced reactions, involving bond breaking and bond making, determine the fate of the excited molecules and it is therefore crucial to understand the mechanism of the initial reaction steps leading to the functional transitions. For several decades, ultrafast bond-breaking processes in various molecular systems have been studied intensively using time-resolved techniques^{1–3}. Unlike bond breaking, which is essentially a unimolecular process and can therefore be initiated by laser photolysis in a synchronized manner, bond making is in most cases a bimolecular process that requires two reactant parties to meet each other to form a chemical bond. The reaction rate of a bimolecular process is generally limited by slow diffusion of the reactants through the solvent and it is thus difficult to synchronize laser excitation with the moment that they meet. Therefore, it is challenging to initiate and follow the bimolecular process with ultrahigh time resolution, although a few special experimental schemes have overcome this obstacle in the case of electron or proton transfer reactions^{8,9}.

In this regard, a Au oligomer complex, $[\text{Au}(\text{CN})_2^-]_n$, offers a good model system in which to study the dynamics of bond formation in solution^{4,10–12}. Au(I) atoms in $[\text{Au}(\text{CN})_2^-]_n$ experience a non-covalent interatomic interaction caused by the relativistic effect called *aurophilicity*^{11,12}. Owing to *aurophilicity*, Au(I) atoms can be weakly bound to each other by van der Waals interactions, forming an aggregate complex

$[\text{Au}(\text{CN})_2^-]_n$ even without covalent bonds. On photoexcitation of the complex, an electron is excited from an antibonding orbital to a bonding orbital, leading to the formation of covalent bonds among Au atoms⁴. Because Au atoms in the ground state of $[\text{Au}(\text{CN})_2^-]_n$ are located in close proximity within the same solvent cage, the formation of covalent Au–Au bonds occurs without being limited by slow diffusion through the solvent. Therefore, the ultrahigh time resolution necessary to probe this bond-making process can be achieved, as in typical unimolecular reactions synchronized with laser photolysis, but the ensuing reaction is like a bimolecular reaction between $\text{Au}(\text{CN})_2^-$ monomers.

Recently, ultrafast Au–Au bond formation in a Au trimer complex, $[\text{Au}(\text{CN})_2^-]_3$, was investigated using transient absorption spectroscopy⁵, and the transient changes in absorption were observed with time constants of variously 500 fs, 2 ps and 2 ns. The first kinetic component (500 fs) was ascribed to the intersystem crossing to a triplet state, which is presumably preceded by rapid contraction of Au–Au bonds within 500 fs. The second kinetic component (2 ps) was assigned to conformational change from bent to linear structure. However, because the transient absorption signal is not directly related to the molecular structure, those structural assignments were based solely on theoretical electronic absorption spectra of model structures. As a result, those structural assignments were disputed by a recent study using quantum chemical calculation⁶. According to the *ab initio* molecular dynamics simulation, the bent-to-linear transition occurs on the 500 fs timescale but was assigned to the 2 ps kinetic component in the transient absorption study. Such discrepancies between experiment and theory, mainly due to limited structural information obtained from experiments, are not limited to this particular system but are a common problem in chemistry in general.

To resolve this discrepancy, we applied time-resolved X-ray solution scattering¹³ (TRXSS) to the same Au trimer complex. TRXSS is an effective method for probing photoinduced structural changes of molecules in solution and has been used to reveal the dynamics and mechanism of various molecular reaction systems ranging from small molecules^{14–17} to biological macromolecules^{18,19}. Although subpicosecond time resolution has been achieved in the case of X-ray^{20–23} and electron^{24–26} diffraction of solid samples and in X-ray absorption spectroscopy²⁷ of liquid samples, the temporal resolution of TRXSS based on synchrotron radiation¹³ has been limited to only 100 ps, thus preventing the observation of ultrafast processes on the timescales of femtoseconds to picoseconds. This limit can be overcome with the recent development of X-ray free-electron lasers (XFELs), which generate ultrashort (< 100 fs long) X-ray pulses with $\sim 10^{12}$ photons per pulse^{28,29}. As a result, it has become possible to explore chemical and biological processes occurring on subpicosecond timescales using TRXSS³⁰. In this work, by performing

¹Center for Nanomaterials and Chemical Reactions, Institute for Basic Science, Daejeon 305-701, South Korea. ²Department of Chemistry, KAIST, Daejeon 305-701, South Korea. ³Institute of Materials Structure Science, High Energy Accelerator Research Organization, 1-1 Oho, Tsukuba, Ibaraki 305-0801, Japan. ⁴RIKEN SPring-8 Center, Kouto 1-1-1, Sayo, Hyogo 679-5148, Japan. ⁵Japan Synchrotron Radiation Research Institute, Kouto 1-1-1, Sayo, Hyogo 679-5198, Japan. ⁶Department of Chemistry, The Catholic University of Korea, Bucheon 420-743, South Korea. ⁷Department of Chemistry, Inha University, Incheon 402-751, South Korea. ⁸Department of Materials Structure Science, School of High Energy Accelerator Science, The Graduate University for Advanced Studies, 1-1 Oho, Tsukuba, Ibaraki 305-0801, Japan. [†]Present addresses: Center for Free-Electron Laser Science, Deutsches Elektronen-Synchrotron, Notkestrasse 85, 22607 Hamburg, Germany (Tokushi Sato); Department of Chemistry, School of Science, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan (Takahiro Sato); Japan Atomic Energy Agency, 8-1-7 Umemidai, Kizugawa, Kyoto 619-0215, Japan (K.O.).

*These authors contributed equally to this work.

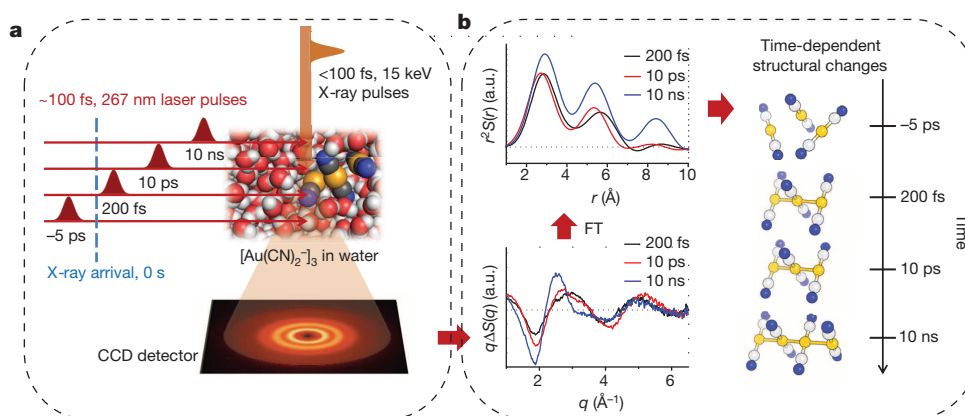


Figure 1 | Femtosecond time-resolved X-ray solution scattering at the XFEL facility and the data analysis. **a**, The photochemical reaction of solutes supplied by a liquid-flowing system is triggered by a femtosecond optical laser pulse. Subsequently, a time-delayed X-ray pulse synchronized with the laser pulse probes the structural dynamics of the reaction. The scattering pattern is detected by a fast two-dimensional charge-coupled device (CCD) detector as shown at the bottom. We measure time-resolved scattering patterns while varying the time delay between the laser and X-ray pulses. **b**, By integrating the

two-dimensional scattering pattern azimuthally, subtracting solvent contributions, performing a Fourier transform (FT) and compensating for the depletion of the initial solute contribution due to photochemical reaction, we obtain one-dimensional RDFs in real space as shown in the plot at the top left. These display the interatomic distances of transient species and products. In this way, Au–Au bond lengths of the $[\text{Au}(\text{CN})_2]^-$ complex can be identified with sub-ångström accuracy, and the time-dependent structural changes of the metal complex can be determined in real time.

the TRXSS experiment at an XFEL facility⁷ (SACLA), we were able to study the ultrafast structural dynamics of bond formation in $[\text{Au}(\text{CN})_2]^-$ in solution with subpicosecond time resolution and sub-ångström spatial resolution.

The femtosecond TRXSS experiment performed in this study is shown schematically in Fig. 1, and the details of the experimental procedure and the analysis are described in Methods and Extended Data Table 1.

For the sample, we used an aqueous solution of $\text{Au}(\text{CN})_2^-$ at a concentration of 300 mM, because in a solution of this concentration the $[\text{Au}(\text{CN})_2]^-$ trimer has much higher absorbance at 267 nm than does a monomer or a dimer of $\text{Au}(\text{CN})_2^-$ (Supplementary Information). As a result, the laser excitation at 267 nm used in our TRXSS experiment predominantly excites the $[\text{Au}(\text{CN})_2]^-$ trimer while exciting the other species negligibly. From the TRXSS experiment, we obtained difference

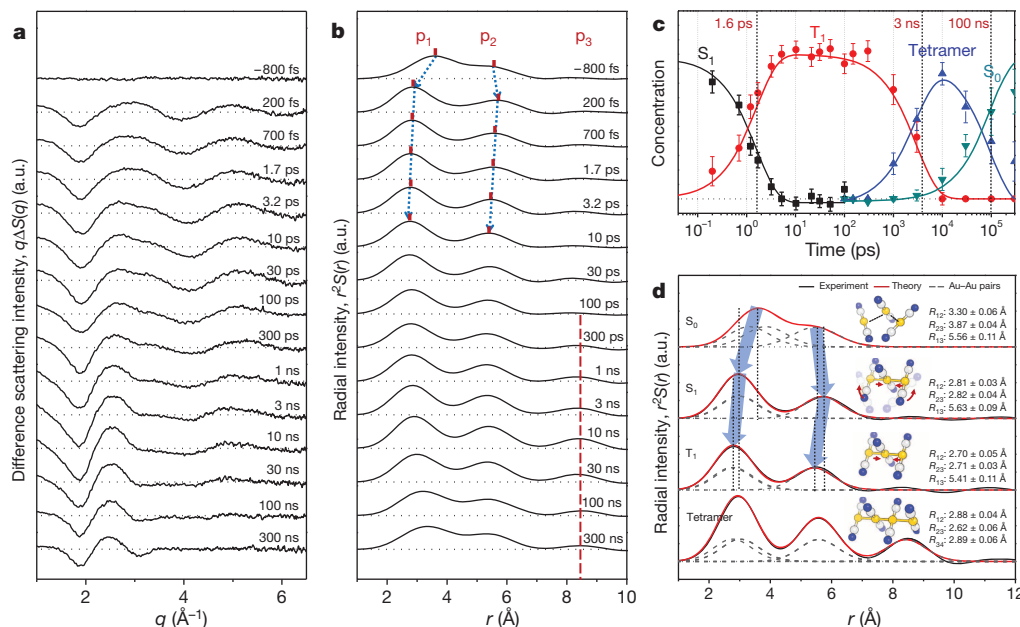


Figure 2 | Time-dependent structural changes of $[\text{Au}(\text{CN})_2]^-$.

a, Experimental difference scattering curves, $q\Delta S(q)$, measured at various time delays from -800 fs to 300 ns (black). For clarity, only data at selected time delays are shown. a.u., arbitrary units. **b**, RDFs, $r^2S(r)$, obtained by Fourier sine transformation of $q\Delta S(q)$ after subtracting solvent contributions. The RDF of the S_0 state was added to the RDFs at all time delays to emphasize only the contributions of the transient solute species associated with bond formation. The blue dashed arrows indicate the time-dependent changes in the locations of the p_1 and p_2 peaks. The red dashed line represents the position of the p_3 peak, corresponding to the signature of the tetramer. **c**, Time-dependent concentrations of the four states and their transition kinetics. The notation for each species is indicated above each trace. The error bar at each data point

indicates the standard error determined from 50 independent measurements (that is, 50 scattering images). The vertical black dotted lines indicate the temporal positions corresponding to the time constants of the three kinetic components. **d**, Species-associated RDFs of the four structures obtained from the singular value decomposition and principal-component analyses (black) and their fits (red) obtained by using model structures containing multiple Au–Au pairs. The blue arrows indicate the changes in R_{12} , R_{23} and R_{13} as transitions occur between states. As fitting parameters, we considered three Au–Au pairs for the S_0 , S_1 and T_1 states and six Au–Au pairs for the tetramer. For each state, the structural parameters obtained from the fits are shown along with their standard errors determined from 50 independent measurements.

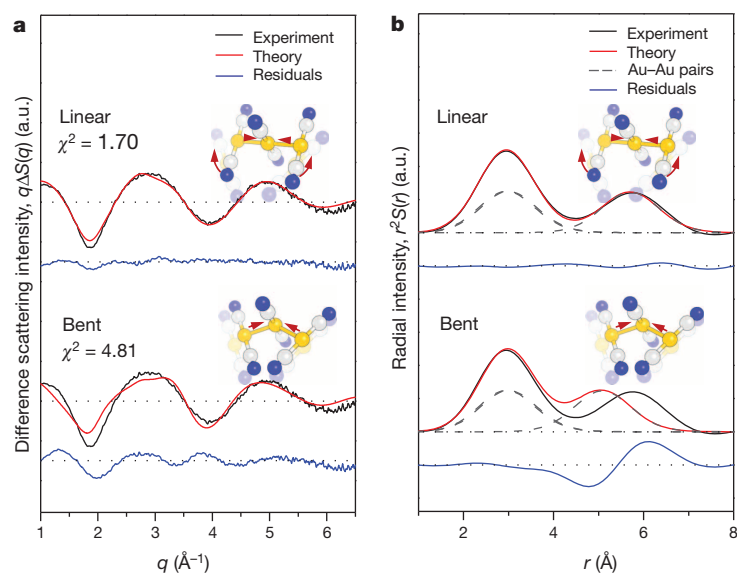


Figure 3 | Structure determination of the S_1 state using the experimental scattering curve at 200 fs time delay in momentum space and real space. **a**, Theoretical difference scattering curves (red) for linear (upper) and bent (lower) structures shown together with the experimental difference scattering curve at 200 fs (black). The residuals (blue) between the theoretical and the experimental curves are shown together. The linear structure gives a much

better fit than the bent structure, which has the same Au–Au–Au bond angle as the S_0 state, thereby indicating that the bent-to-linear transition is completed at 200 fs time delay. **b**, Corresponding experimental (black) and theoretical (red) RDFs, $rS(r)$. It can be seen that in the bent structure R_{13} is too small to fit the experimental RDF at 200 fs.

scattering curves, $q\Delta S(q, t)$, where q is the momentum transfer between the incident and the elastically scattered X-ray waves and ΔS is the difference between scattering intensities measured before the laser excitation (that is, at a negative time delay) and after the laser excitation (that is, at a positive time delay). The experimental difference scattering curves

measured at various time delays from –800 fs to 300 ns are shown in Fig. 2a. The difference scattering curves show distinct oscillatory features along the q axis, which indicate large structural changes in $[\text{Au}(\text{CN})_2]_3^-$ during the formation of covalent Au–Au bonds. Considering that the oscillatory features appear distinct even after the shortest time delay

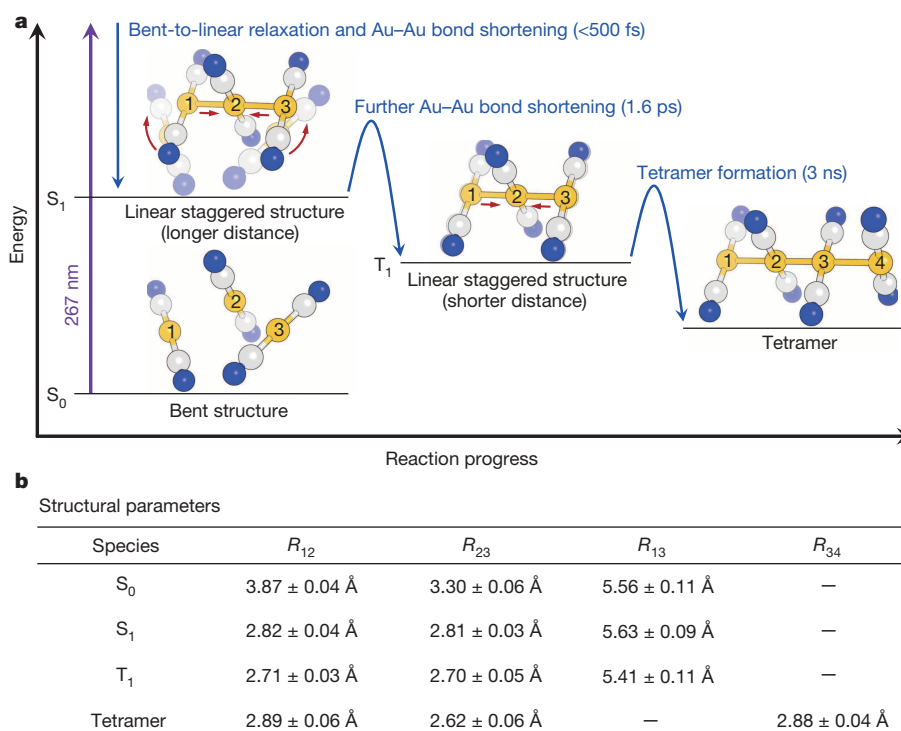


Figure 4 | Mechanism of photoinduced bond formation in $[\text{Au}(\text{CN})_2]_3^-$. **a**, Femtosecond TRXSS reveals the dynamics and the atomic movements associated with the Au–Au bond formation in real time with sub-ångström spatial resolution. The S_0 state with weakly bound Au atoms in a bent geometry transforms to the S_1 state with tightly bound Au atoms in a linear geometry.

Subsequently, the S_1 state transforms first to the T_1 state, with further contraction of Au–Au bonds, and then to a tetramer through formation of another Au–Au bond. **b**, Structural parameters of each state and their standard errors determined from 50 independent measurements.

(200 fs), the first step of bond formation in $[\text{Au}(\text{CN})_2]^-_3$ must occur impulsively within the time resolution (~ 500 fs). The oscillatory features change further over time until only the oscillatory contribution from solvent heating remains after 300 ns (Extended Data Fig. 2).

A more intuitive picture of the structural change in $[\text{Au}(\text{CN})_2]^-_3$ can be obtained when we convert $q\Delta S(q, t)$ into difference radial distribution functions (RDFs), $r^2\Delta S(r, t)$, in real space by Fourier sine transformation. To emphasize only the contributions of transient solute species associated with bond formation, we added the RDF of the ground (S_0) state to the difference RDFs at all time delays and obtained the RDFs $r^2S(r, t)$ shown in Fig. 2b. Because the solvent contributions were eliminated and the contributions from C and N atoms in $[\text{Au}(\text{CN})_2]^-_3$ are almost negligible compared with the strong scattering from Au atoms (Supplementary Information and Extended Data Fig. 6), the RDFs shown in Fig. 2b actually represent the interatomic distances among Au atoms of $[\text{Au}(\text{CN})_2]^-_3$ in real space.

It can be seen that two peaks (p_1 and p_2) are distinct in the RDFs in Fig. 2b. Because $[\text{Au}(\text{CN})_2]^-_3$ is a trimer, we can assign the p_1 peak to the $\text{Au}_1\text{--Au}_2$ pair (bond length, R_{12}) and the $\text{Au}_2\text{--Au}_3$ pair (bond length, R_{23}), and the p_2 peak to the $\text{Au}_1\text{--Au}_3$ pair (bond length, R_{13}). This assignment is supported by the observation that the intensity of p_1 is about twice as large as that of p_2 . In the S_0 state of $[\text{Au}(\text{CN})_2]^-_3$ (that is, RDF at ~ 800 fs time delay), R_{12} and R_{23} (~ 3.6 Å as indicated by the position of p_1) are relatively large compared with the typical length of an Au–Au covalent bond (~ 2.7 Å), indicating that the Au atoms are weakly bound, and R_{13} (~ 5.56 Å as indicated by the position of p_2) is smaller than the sum of R_{12} and R_{23} , indicating that the S_0 state has a bent structure. As expected in the momentum-space data, the RDF at 200 fs time delay is significantly different from the one at ~ 800 fs, suggesting that the first step of bond formation in $[\text{Au}(\text{CN})_2]^-_3$ occurs within the time resolution of our experiment. Compared with the S_0 state, at 200 fs time delay R_{13} has increased slightly and R_{12} and R_{23} have decreased significantly, indicating the formation of covalent Au–Au bonds at this step. Simultaneously, R_{13} (~ 5.63 Å) becomes similar to the sum of R_{12} and R_{23} (~ 2.8 Å each), which is evidence of a conformational transition from bent to linear geometry. We note that the timescale of the bent-to-linear transition determined from our TRXSS experiment is in good agreement with the timescale predicted from the previous theoretical study⁶. From 200 fs to 10 ps, the p_1 and p_2 peaks shift to smaller distances, indicating further decrease of the Au–Au distances due to the formation of stronger covalent Au–Au bonds. In this time range, the ratio between $R_{12} + R_{23}$ and R_{13} remains 1:1, showing that the linear structure is preserved. After 100 ps, the p_3 peak appears at ~ 8.5 Å and increases until a time delay of 10 ns. Because the position of the p_3 peak corresponds to too great a Au–Au distance for the complex to be a trimer, p_3 must be a signature of the formation of a tetramer complex. Also, increased intensities of the p_1 and p_2 peaks imply the presence of a larger number of Au–Au pairs in the tetramer. After 10 ns, the RDF returns gradually to the RDF of the S_0 state.

By singular value decomposition (SVD) and principal-component analysis, we obtained species-associated RDFs for four states, the ground (S_0) state, an excited (S_1) state, a triplet (T_1) state and a tetramer (Fig. 2d), as well as their kinetics. We fitted the experimental RDFs at various time delays by linear combinations of the species-associated RDFs, and determined the time-dependent concentration of each state from the coefficient of the corresponding species-associated RDF (Fig. 2c). As a result, we obtained three kinetic components with time constants of 1.6 ± 0.1 ps, 3 ± 0.5 ns and 100 ± 20 ns, which correspond respectively to the transition from S_1 to T_1 , the transition from T_1 to the tetramer, and the transition from the tetramer to S_0 . The timescales of the three kinetic components match well with the ones identified in the previous transient absorption study⁵, except our TRXSS data lack the ~ 500 fs component, which in the transient absorption study was assigned to intersystem crossing to a triplet state. Here we note that TRXSS is sensitive only to processes accompanying structural changes. Therefore, the fact that the ~ 500 fs kinetic component is not observed by TRXSS

indicates that intersystem crossing does not involve any significant structural change (see Extended Data Fig. 9 for a detailed kinetic scheme).

To reconstruct the structures of the four states (S_0 , S_1 , T_1 and tetramer) and extract the Au–Au distances for each state, we performed a structural fitting analysis for the species-associated RDF of each state using the Au–Au distances as fitting parameters. The fitting results for the four states are shown in Fig. 2d. Reconstructed structures based on the optimized Au–Au distances are shown together. The reconstructed structures of the four states presented in Fig. 2d are in good agreement with the structural changes inferred from Fig. 2b. The S_0 state has weakly bound Au atoms in a bent geometry with R_{12} and R_{23} different from each other, which difference can be attributed to a broadening of the RDF induced by relatively free movements of the weakly bound Au atoms (Supplementary Information and Extended Data Fig. 10). The S_1 state has much shorter Au–Au distances than does the S_0 state, owing to the formation of covalent Au–Au bonds. Notably, the S_1 state has Au atoms aligned in a linear geometry, confirming that the bent-to-linear transition occurs during the transition to the S_1 state. We also find that the two Au–Au bonds in the S_1 state are the same length, indicating the symmetric structure of S_1 . The T_1 state has even shorter Au–Au distances than does S_1 , owing to the formation of stronger covalent Au–Au bonds, and retains a linear and symmetric structure. Finally, we clearly identified the structure of the tetramer, $[\text{Au}(\text{CN})_2]^-_4$, as the final species formed before ultimately returning to the S_0 state. The changes in the Au–Au distances and the conformations of the S_0 , S_1 , T_1 and tetramer states in Fig. 2d are in good agreement with the results of the previous theoretical calculation⁶.

The previous transient absorption study identified the changes of transient absorption on 500 fs, 2 ps and 2 ns timescales⁵. The timescales of the transitions found in the transient absorption study match well the results of this work, but the detailed structural changes were assigned differently. In particular, from our TRXSS measurement we inferred that the bent-to-linear transition occurs within a few hundred femtoseconds rather than on the timescale of 2 ps. To account for this discrepancy, we determined the structure of the S_1 state more carefully by fitting the experimental difference scattering curve at 200 fs time delay using two different model structures (Fig. 3). One is a linear structure where R_{13} is equal to the sum of R_{12} and R_{23} , and the other is a bent structure where the Au–Au–Au bond angle of the S_1 state is the same as that of the S_0 state. In Fig. 3a, it is clearly seen that the theoretical scattering curve calculated from the linear structure ($\chi^2 = 1.70$) fits the experimental scattering curve at 200 fs time delay much better than does the one calculated from the bent structure ($\chi^2 = 4.81$), where χ^2 is a quantitative measure of the discrepancy between the experimental and theoretical scattering curves (Methods). The difference between the linear and bent structures can be seen more distinctly in the real-space RDFs in Fig. 3b. In the bent structure, R_{13} is too small to fit the experimental RDF at 200 fs. Therefore, the bent-to-linear transition must be completed at 200 fs time delay.

On the basis of the reconstructed structures of the four different states and the transition dynamics among them, in Fig. 4 we summarize the mechanism for photoinduced formation of Au–Au covalent bonds in $[\text{Au}(\text{CN})_2]^-_3$. The S_0 state with weakly bound Au atoms in a bent geometry transforms within a few hundred femtoseconds to the S_1 state with tightly bound Au atoms (2.8 Å Au–Au distances) in a linear and symmetric geometry. The S_1 state transforms to the T_1 state with a 1.6 ps time constant accompanying further contraction of the Au–Au bonds by 0.1 Å. Then the T_1 state converts to a tetramer in 3 ns through formation of another Au–Au bond, and the S_0 state is ultimately recovered in ~ 100 ns.

We have demonstrated the capability of XFEL-based femtosecond TRXSS by both determining the overall mechanism for the formation of covalent Au–Au bonds in the $[\text{Au}(\text{CN})_2]^-_3$ complex and gaining rich structural information. Femtosecond TRXSS offers a means of visualizing the entire process of photoinduced reactions in real time and real space, and can be used as a fundamental tool to study the reaction dynamics of chemical and biological systems.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 5 September; accepted 19 December 2014.

- Zewail, A. H. Laser femtochemistry. *Science* **242**, 1645–1653 (1988).
- Harris, A. L., Brown, J. K. & Harris, C. B. The nature of simple photodissociation reactions in liquids on ultrafast time scales. *Annu. Rev. Phys. Chem.* **39**, 341–366 (1988).
- Jonas, D. M., Bradforth, S. E., Passino, S. A. & Fleming, G. R. Femtosecond wavepacket spectroscopy: influence of temperature, wavelength, and pulse duration. *J. Phys. Chem.* **99**, 2594–2608 (1995).
- Rawashdeh-Omary, M. A., Omary, M. A., Patterson, H. H. & Fackler, J. P. Excited-state interactions for $[\text{Au}(\text{CN})_2]_n$ and $[\text{Ag}(\text{CN})_2]_n$ oligomers in solution. Formation of luminescent gold-gold bonded excimers and exciplexes. *J. Am. Chem. Soc.* **123**, 11237–11247 (2001).
- Iwamura, M., Nozaki, K., Takeuchi, S. & Tahara, T. Real-time observation of tight Au–Au bond formation and relevant coherent motion upon photoexcitation of $[\text{Au}(\text{CN})_2]_n$ oligomers. *J. Am. Chem. Soc.* **135**, 538–541 (2013).
- Cui, G. L., Cao, X. Y., Fang, W. H., Dolg, M. & Thiel, W. Photoinduced gold(I)–gold(I) chemical bonding in dicyanoaurate oligomers. *Angew. Chem. Int. Ed.* **52**, 10281–10285 (2013).
- Tamasaku, K. *et al.* X-ray two-photon absorption competing against single and sequential multiphoton processes. *Nature Photon.* **8**, 313–316 (2014).
- Rini, M., Magnes, B. Z., Pines, E. & Nibbering, E. T. J. Real-time observation of bimodal proton transfer in acid-base pairs in water. *Science* **301**, 349–352 (2003).
- Rosspeintner, A., Lang, B. & Vauthey, E. Ultrafast photochemistry in liquids. *Annu. Rev. Phys. Chem.* **64**, 247–271 (2013).
- Pyykkö, P. Theoretical chemistry of gold. *Angew. Chem. Int. Ed.* **43**, 4412–4456 (2004).
- Wang, S. G. & Schwarz, W. H. E. Quasi-relativistic density functional study of aurophilic interactions. *J. Am. Chem. Soc.* **126**, 1266–1276 (2004).
- Schmidbaur, H. & Schier, A. A briefing on aurophilicity. *Chem. Soc. Rev.* **37**, 1931–1951 (2008).
- Ihee, H. Visualizing solution-phase reaction dynamics with time-resolved X-ray liquidography. *Acc. Chem. Res.* **42**, 356–366 (2009).
- Ihee, H. *et al.* Ultrafast X-ray diffraction of transient molecular structures in solution. *Science* **309**, 1223–1227 (2005).
- Davidsson, J. *et al.* Structural determination of a transient isomer of CH_2I_2 by picosecond x-ray diffraction. *Phys. Rev. Lett.* **94**, 245503 (2005).
- Christensen, M. *et al.* Time-resolved X-ray scattering of an electronically excited state in solution. Structure of the $^3A_{2u}$ state of tetrakis- μ -pyrophosphitodiplatinate(II). *J. Am. Chem. Soc.* **131**, 502–508 (2009).
- Kim, K. H. *et al.* Solvent-dependent molecular structure of ionic species directly measured by ultrafast X-ray solution scattering. *Phys. Rev. Lett.* **110**, 165505 (2013).
- Plech, A., Kotaidis, V., Lorenc, M. & Boneberg, J. Femtosecond laser near-field ablation from gold nanoparticles. *Nature Phys.* **2**, 44–47 (2006).
- Kim, K. H. *et al.* Direct observation of cooperative protein structural dynamics of homodimeric hemoglobin from 100 ps to 10 ms with pump-probe X-ray solution scattering. *J. Am. Chem. Soc.* **134**, 7001–7008 (2012).
- Sokolowski-Tinten, K. *et al.* Femtosecond X-ray measurement of coherent lattice vibrations near the Lindemann stability limit. *Nature* **422**, 287–289 (2003).
- Fritz, D. M. *et al.* Ultrafast bond softening in bismuth: mapping a solid's interatomic potential with X-rays. *Science* **315**, 633–636 (2007).
- Coppens, P. Molecular excited-state structure by time-resolved pump-probe X-ray diffraction. What is new and what are the prospects for further progress? *J. Phys. Chem. Lett.* **2**, 616–621 (2011).
- Miller, T. A. *et al.* The mechanism of ultrafast structural switching in superionic copper(I) sulphide nanocrystals. *Nature Commun.* **4**, 1369 (2013).
- Zewail, A. H. Four-dimensional electron microscopy. *Science* **328**, 187–193 (2010).
- Kirchner, F. O., Gliserin, A., Krausz, F. & Baum, P. Laser streaking of free electrons at 25 keV. *Nature Photon.* **8**, 52–57 (2014).
- Miller, R. J. D. Mapping atomic motions with ultrabright electrons: the chemists' gedanken experiment enters the lab frame. *Annu. Rev. Phys. Chem.* **65**, 583–604 (2014).
- Bressler, C. *et al.* Femtosecond XANES study of the light-induced spin crossover dynamics in an iron(II) complex. *Science* **323**, 489–492 (2009).
- Lemke, H. T. *et al.* Femtosecond X-ray absorption spectroscopy at a hard X-ray free electron laser: application to spin crossover dynamics. *J. Phys. Chem. A* **117**, 735–740 (2013).
- Zhang, W. *et al.* Tracking excited-state charge and spin dynamics in iron coordination complexes. *Nature* **509**, 345–348 (2014).
- Arnlund, D. *et al.* Visualizing a protein quake with time-resolved X-ray scattering at a free-electron laser. *Nature Methods* **11**, 923–926 (2014).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank M. Iwamura and K. Nozaki for discussions. This work was supported by IBS-R004-G2; the X-ray Free Electron Laser Priority Strategic Program of MEXT, Japan; PRESTO/JST; the Innovative Areas 'Artificial Photosynthesis (AnApple)' (no. 25107527) grant from the Japan Society for the Promotion of Science; and the Basic Science Research Program through the National Research Foundation of Korea, funded by the Ministry of Science, ICT & Future Planning (NRF-2014R1A1A1002511). The experiments were performed at beamline BL3 of SACLA with the approval of the Japan Synchrotron Radiation Research Institute (proposal nos 2012A8030, 2012A8038, 2012B8029, 2012B8043, 2013A8053, 2013B8036, 2013B8059, 2014A8042 and 2014A8022) and at beamline NW14A of KEK with the approval of the Photon Factory Program Advisory Committee (proposal nos 2011G655, 2012G778 and 2012G779).

Author Contributions H.I. and S.-i.A. designed the study. K.H.K., J.G.K., S.N., Tokushi Sato, K.Y.O., T.W.K., H.K., J.J., S.P., C.S., Takahiro Sato, K.O., T.T., K.T., M.Y., T.I., Jeongho Kim, H.I. and S.-i.A. did the experiment. K.H.K., J.G.K., S.N., Tokushi Sato and Joonghan Kim analysed the data. K.H.K., J.G.K., S.N., K.Y.O., R.R., Jeongho Kim, H.I. and S.-i.A. wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to H.I. (hyotcherl.ihee@kaist.ac.kr) or S.-i.A. (shinichi.adachi@kek.jp).

METHODS

TRXSS data collection. The experimental set-up of femtosecond X-ray solution scattering is shown schematically in Fig. 1 in the main text. A laser pulse initiates a photoinduced reaction of the sample molecules and a time-delayed X-ray pulse probes the progress of the reaction. Time-resolved X-ray solution scattering data were collected at the BL3 beamline of SACLA and the NW14A beamline of KEK. X-ray pulses with sub-100 fs duration generated from SACLA were used for measuring the data at early time delays (from -800 fs to 100 ps), while X-ray pulses with 100 ps duration generated from KEK were used for measuring the data at late time delays (from 100 ps to 1 μ s). To check the reproducibility of the X-ray scattering signals at the two beamlines, we compared time-resolved difference scattering curves, $q\Delta S(q, t)$, measured at SACLA and KEK as shown in Extended Data Fig. 1a. The two difference scattering curves at a common time delay (100 ps) are identical to each other within the experimental error, indicating that our measurement is highly reproducible and independent of the facility. Difference scattering curves in the entire time range (from -800 fs to 1 μ s) are shown in Extended Data Fig. 1b.

TRXSS data collection at SACLA. Time-resolved X-ray solution scattering measurements at early time delays (from -800 fs to 100 ps) were performed at the BL3 beamline of SACLA. Femtosecond laser pulses at 800 nm centre wavelength were generated from the Ti:sapphire regenerative amplifier and converted to 100 fs pulses at 267 nm wavelength by third-harmonic generation. The laser beam was focused by a lens to a spot of 300 μ m diameter, where the laser beam was overlapped with the X-ray beam with the crossing angle of 10° . The femtosecond X-ray pulses were generated from the XFEL at SACLA by self-amplified spontaneous emission (SASE). The X-ray pulses have a centre energy of 15 keV and an energy bandwidth narrow enough ($\Delta E/E = 0.6\%$) for monitoring photoinduced structural changes of small molecules^{31,32}. The X-ray beam was focused on a spot of 200 μ m diameter at the sample position and the resultant X-ray fluence was 1.3 mJ mm^{-2} . The scattering patterns generated by X-ray pulses were measured with an area detector (Rayonix LX255-HS) with a sample-to-detector distance of 31 mm. We used aqueous solution of the oligomer of Au complex $[\text{Au}(\text{CN})_2]_n$. The aggregation number of the Au oligomer and the position of an absorption peak in the ultraviolet region change depending on the concentration of the sample solution³³. In this work, we focus on the trimer, $[\text{Au}(\text{CN})_2]_3$, that is formed in the solution of 300 mM concentration and can be excited by laser pulses of 267 nm centre wavelength. As shown in the Extended Data Fig. 7 and Supplementary Information, the signal is dominated by the contribution of the trimer while the contributions of other species are negligible. In fact, as shown in Extended Data Fig. 8, the TRXSS signals measured with the excitation at 267 and 310 nm, respectively, are identical to each other in terms of the kinetics and the shape of the difference scattering curves, confirming the dominant contribution of the trimer to the TRXSS signal. On photoexcitation, the trimer undergoes structural changes including variation in its geometric structure and the formation of covalent Au–Au bonds. The sample solution was circulated through a sapphire nozzle with a 100 μ m-thick aperture. To supply fresh sample for every laser and X-ray shot, the sample-flowing velocity was set to be over 3 m s^{-1} . To prevent the scattering signal from being contaminated by radiation-damaged sample, the sample in the reservoir was replaced with fresh sample whenever the reservoir sample failed to produce the transient signal measured at 100 ps. Even if the transient signal at 100 ps did not change, the sample in the reservoir was replaced with fresh one regularly (every 2–3 h of measurement) to ensure the delivery of fresh sample. The structural change was monitored by the scattering patterns generated by X-ray pulses of sub-100 fs duration. The time resolution of the X-ray solution scattering experiment was $\sim 500 \text{ fs}$, which was limited by the timing jitter between the laser and X-ray pulses as well as a velocity mismatch of 120 fs that was calculated by considering the laser/X-ray crossing angle of 10° and the sample thickness of 100 μ m. The laser-off images were acquired with the X-ray pulse arriving 5 ps earlier than the laser pulse (that is, -5 ps time delay), to probe the (unexcited) molecules in the ground state while assuring the same average temperature of the sample solution. These laser-off images were repeatedly measured before every laser-on image and were used as a reference for calculating the time-resolved difference X-ray scattering patterns. To achieve a signal-to-noise ratio high enough for data analysis, about 50 images were acquired at each time delay. Each scattering image was obtained by accumulating scattering intensities of 80 X-ray pulses. The data collection scheme of accumulating multiple X-ray shots in a single scattering image was enough to effectively suppress the fluctuation of signal caused by random pulse energies of the SASE process at SACLA, as proven by the good agreement of two difference scattering curves measured at the XFEL and synchrotron radiation source (Extended Data Fig. 1a). The scattering curves were measured at the following time delays: -5 ps, -800 fs, -300 fs, 200 fs, 700 fs, 1.2 ps, 1.7 ps, 3.2 ps, 5.2 ps, 10 ps, 20 ps, 30 ps, 50 ps, 100 ps. The scattering signals arising from solvent (water) heating were also measured using 40 mM FeCl_3 solution with the same experimental conditions. Details of the data collection parameters are summarized in Supplementary Table 1.

TRXSS data collection at KEK. Time-resolved X-ray solution scattering measurement at late time delays (100 ps–1 μ s) was performed at the NW14A beamline of KEK by following the experimental protocol described in our previous publications^{27,34,35}. Third-harmonic generation of the 800 nm output pulses from an amplified Ti:sapphire laser system provided 150 fs pulses at 267 nm centre wavelength. The laser beam was focused by a lens to a spot of 300 μ m diameter, where the laser beam is overlapped with the X-ray beam with a crossing angle of 10° . The laser pulses were synchronized with X-ray pulses from the synchrotron using an active feedback control loop that adjusts the laser oscillator cavity length, and the relative time delay between the laser and X-ray pulses was controlled electronically. The time-delayed X-ray pulses were selected by using a synchronized mechanical chopper. A multilayer optic coated with depth-graded Ru/C layers ($d = 40 \text{ \AA}$, NTT Advanced Technology) produced a Gaussian-type X-ray spectrum with a centre energy of 15.6 keV and a $\sim 5\%$ energy bandwidth. The X-ray beam was focused on a spot of 200 μ m diameter at the sample position, and the resultant X-ray fluence was 0.017 mJ mm^{-2} . The scattering patterns generated by X-ray pulses of 100 ps (full-width at half-maximum) duration were measured with an area detector (MarCCD165, Mar USA) with a sample-to-detector distance of 40 mm. The sample solution of the same concentration as used at SACLA was circulated through a sapphire nozzle with a 300 μ m-thick aperture. The sample flow velocity was set to be over 3 m s^{-1} to supply fresh sample for every X-ray and laser shot. The sample in the reservoir was replaced with fresh sample whenever the reservoir sample failed to produce the transient signal measured at 100 ps. Even if the transient signal at 100 ps did not change, the sample in the reservoir was replaced with fresh one regularly (every 2–3 h of measurement) to ensure the delivery of fresh sample. The time resolution of the X-ray solution scattering experiment was 100 ps, which was limited by the duration of the X-ray pulses. The laser-off images were acquired with the X-ray pulse arriving 3 ns earlier than the laser pulse (that is, -3 ns time delay), to eliminate the contribution of the (unexcited) ground-state reactants. These laser-off images were used as a reference for calculating the time-resolved difference X-ray scattering patterns. To achieve a signal-to-noise ratio high enough for data analysis, more than 50 images were acquired and averaged at each time delay. The scattering curves were measured at the following time delays: -3 ns, -150 ps, 100 ps, 150 ps, 300 ps, 1 ns, 3 ns, 10 ns, 30 ns, 100 ns, 300 ns and 1 μ s. The scattering signals arising from solvent (water) heating were also measured using 40 mM FeCl_3 solution with the same experimental conditions. Details of the data collection parameters are summarized in Supplementary Table 1.

Removal of the solvent contribution. To study only the dynamics of the Au–Au bond formation, the scattering arising from heating of pure solvent induced by laser excitation needs to be subtracted from the experimental scattering data. For this purpose, a separate time-resolved X-ray scattering experiment was performed on FeCl_3 solution in water. The resultant difference scattering curves are shown in Extended Data Fig. 2a. As can be seen in Extended Data Fig. 2b, c, SVD of the data identifies only one singular component, implying that only a single difference scattering curve accounts for the contribution of solvent heating in the time range up to 100 ps. In addition, as can be seen in Extended Data Fig. 2d, the difference scattering curve of the $[\text{Au}(\text{CN})_2]_3$ solution at 1 μ s time delay is identical to the one for solvent heating, confirming that the difference scattering curves of $[\text{Au}(\text{CN})_2]_3$ at late time delays are dominated by the solvent heating contribution. The amount of heat dissipated in the sample solution can be determined from the scaling between these two curves. The obtained solvent heating contribution was subtracted from the experimental difference scattering curves at all time delays.

Fourier sine transformation of $q\Delta S(q)$. The difference RDF, $r^2\Delta S(r, t)$, is a measure of the radial electron density change as a function of interatomic distance r in real space, and was obtained by Fourier sine transformation of the $q\Delta S(q, t)$ curves:

$$r^2\Delta S(r, t) = \frac{r}{2\pi^2} \int_0^\infty q\Delta S(q, t) \sin(qr) e^{-q^2\alpha} dq$$

Here the exponential with constant $\alpha = 0.03 \text{ \AA}^2$ is a damping term that accounts for the finite q range in the experiment. The resultant $r^2\Delta S(r, t)$ curves are shown in Extended Data Fig. 3.

Singular value decomposition. To determine the kinetic model of the photoinduced reaction of $[\text{Au}(\text{CN})_2]_3$ and obtain the species-associated RDFs for each transient state, we applied the SVD analysis and kinetic analysis to our experimental RDFs. From the time-resolved X-ray scattering data, we can build an $n_r \times n_t$ data matrix, A , where n_r is the number of r points in the RDFs and n_t is the number of time-delay points. The matrix A can be decomposed into three matrices while satisfying the relationship $A = USV^T$, where U is an $n_r \times n_r$ matrix whose columns are called left singular vectors (that is, time-independent r -spectra) of A , V is an $n_t \times n_t$ matrix whose columns are called right singular vectors (that is, amplitude changes in the left singular vectors in U as time evolves) of A , and S is a diagonal $n_r \times n_t$ matrix whose diagonal elements are called singular values of A and can possess only non-negative values. The matrices U and V obey the relationships $U^T U = I_{n_r}$,

and $V^T V = I_{n_t}$, respectively, where I_{n_t} is the $n_t \times n_t$ identity matrix. The diagonal elements of S (that is, singular values) represent the weights of left singular vectors in U . Because the singular values are ordered so that $s_1 \geq s_2 \geq \dots \geq s_n \geq 0$ (both left and right), singular vectors on the left side of the matrix U are supposed to have larger contributions to the experimental TRXSS signal than are the ones on the right side of U . The left singular vectors, when linearly combined together, give information on the RDFs associated with distinct transient species, whereas the right singular vectors contain the information on the population dynamics of the transient species. Thus, the SVD analysis provides a model-independent estimation of the number of structurally distinguishable transient species and the dynamics of each species.

By performing the SVD analysis on our experimental difference RDFs, $r^2 \Delta S(r, t)$, we identified four singular components with significant singular values, indicating the existence of four structurally distinguishable transient states. The right singular vectors of these four significant singular components were fitted by a convolution of a Gaussian function representing the instrument response function (IRF) and a sum of three exponential functions representing transitions among the transient intermediate states. As a result, we obtained exponentials with time constants of 1.6 ± 0.1 ps, 3 ± 0.5 ns and 100 ± 20 ns and an IRF with a 480 ± 10 fs full-width at half-maximum. Thus, we identified four transient states and three kinetic components connecting the four species.

Kinetic analysis. To obtain species-associated RDFs of the four transient states identified in the SVD analysis, we performed kinetic analysis on the U and V matrices using an appropriate kinetic model. First, following the result of the SVD analysis, we defined new matrices, U' , V' and S' , that contain only the first four elements of U , V and S . In other words, U' is an $n_r \times 4$ matrix containing only the first four left singular vectors of U , S' is a 4×4 diagonal matrix containing only the first four singular values of S , and V' is an $n_t \times 4$ matrix containing the first four right singular vectors of V . Among various kinetic models, the only one that can account for the four transient states and the three kinetic components is the sequential model. Therefore, by solving rate equations based on the sequential model, the concentrations of the four transient states can be calculated using the three kinetic components determined from the SVD analysis. We defined a matrix C that represents the time-dependent concentrations of the four transient states and related it to V' using a parameter matrix P that satisfies the relation $V' = CP$. In other words, the linear combination (via P) of the concentrations of the four transient states (C) gives the four right singular vectors constituting V' . In our analysis, C is an $n_t \times 4$ matrix containing the time-dependent concentrations of the four transient states (S_0 , S_1 , T_1 and the tetramer), and P is a 4×4 matrix containing coefficients that relate the time-dependent concentrations of the transient states in C to the right singular vectors in V' . Once we determine C by solving the rate equations and convolving with the IRF, the theoretical RDFs at various time delays, A' , can be generated as follows:

$$\begin{aligned} A' &= U' S' V'^T = U' S' (CP)^T \\ &= U' S' (P^T C^T) = (U' S' P^T) C^T \end{aligned} \quad (1)$$

The matrix P can be optimized by minimizing the discrepancy, χ^2 , between the theoretical and experimental difference scattering curves using the MINUIT package:

$$\chi^2 = \sum_{i=1}^{n_r} \sum_{j=1}^{n_t} \left(\frac{r^2 \Delta S_{\text{exp}}(r_i, t_j) - r^2 \Delta S_{\text{theory}}(r_i, t_j)}{\sigma_{ij}} \right)^2$$

Here $r^2 \Delta S_{\text{exp}}(r_i, t_j)$ and $r^2 \Delta S_{\text{theory}}(r_i, t_j)$ are the experimental and theoretical RDFs at given r and t values, respectively, and σ_{ij} is the experimental standard deviation at given r and t values. From equation (1), we can define a matrix B as $B = U' S' P^T$, that is, a linear combination (via P) of the four left singular vectors constituting U' that are weighted by their singular values in S' . As a result, B , an $n_r \times 4$ matrix, contains the RDFs directly associated with the transient states. Thus, by optimizing P , we obtain the time-independent, species-associated RDFs of the intermediate species (optimized B).

Calculation of theoretical RDFs. For individual transient states, theoretical RDFs were expressed as a sum of multiple RDFs, S_{R_i} , each of which corresponds to an Au–Au pair:

$$r^2 S_{\text{theory}}(r) = r^2 \sum_{i=1}^n S_{R_i}(r) \quad (2)$$

Here R_i is the Au–Au distance for the i th pair of Au atoms. For the trimer states (S_0 , S_1 and T_1) and the tetramer, n was set to be 3 and 6, respectively. Each $S_{R_i}(r)$ curve was calculated by Fourier sine transformation of the theoretical scattering curve, $S_{R_i}(q)$, as follows:

$$r^2 S_{R_i}(r) = \frac{r}{2\pi^2} \int_0^\infty q S_{R_i}(q) \sin(qr) e^{-q^2 z} dq \quad (3)$$

The damping constant (z) used to obtain the experimental RDFs was also used to obtain the theoretical RDFs. The theoretical scattering curves $S_{R_i}(q)$ from Au–Au pairs were obtained by the simple Debye formula

$$S_{R_i}(q) = F_{\text{Au}}^2(q) \frac{\sin(qR_i)}{qR_i} \quad (4)$$

where F_{Au} is the atomic form factor of the Au atom. By substituting equations (3) and (4) into equation (2), we obtained

$$r^2 S_{\text{theory}}(r) = \sum_{i=1}^n \frac{r}{2\pi^2} \int_0^\infty q F_{\text{Au}}^2(q) \frac{\sin(qR_i)}{qR_i} \sin(qr) e^{-q^2 z} dq \quad (5)$$

The theoretical RDFs for the transient states were calculated by equation (5). We note that the only variables in equation (5) are the Au–Au distances.

Structural fitting analysis. To reconstruct the structures of the four states (S_0 , S_1 , T_1 and the tetramer) and extract the Au–Au distances for each state, we performed a structural fitting analysis of the species-associated RDFs of the four states. As fitting parameters of the analysis, we considered three Au–Au distances for the S_0 , S_1 and T_1 states, six Au–Au distances for the tetramer, and a scaling factor between the number of excited molecules and the signal intensity. The maximum-likelihood estimation with the χ^2 estimator^{15,36,37} was used with four (for S_0 , S_1 and T_1) or seven (for the tetramer) variable parameters. The χ^2 estimator is given by

$$\chi^2(R_1, R_2, R_3, A) = \frac{1}{N-p-1} \sum_i \frac{(S_{\text{theory}}(r_i) - S_{\text{exp}}(r_i))^2}{\sigma_i^2}$$

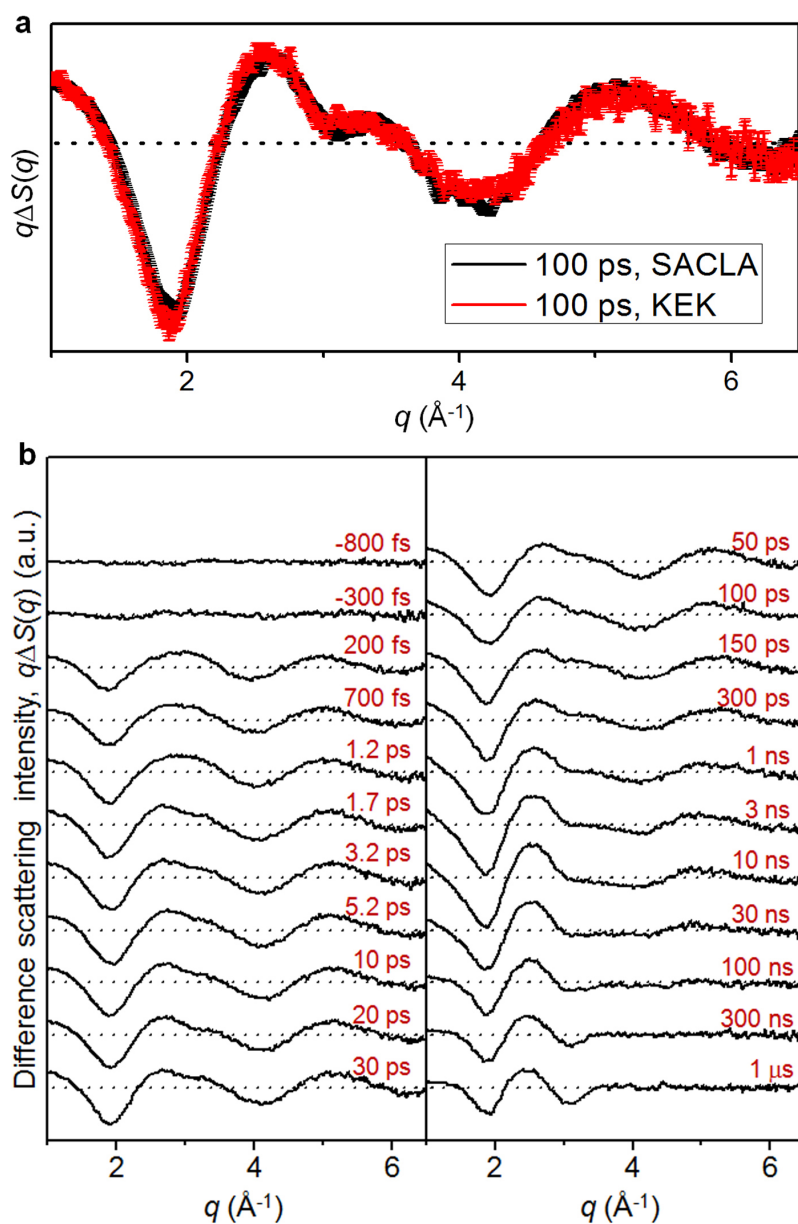
where $N = 500$ is the total number of r points, $p = 4$ or 7 is the number of fitting parameters and σ_i is the standard deviation. The likelihood (L) is related to χ^2 by the following equation:

$$L(R_1, R_2, R_3, A) \propto \exp(-\chi^2/2)$$

The errors of multiple fitting parameters can be determined from this relationship by calculating the boundary values at 68.3% of the likelihood distribution. The calculation was performed by the MINUIT software package, and the error values were provided by MINOS algorithm in MINUIT. Because we used the standard deviation of the measurement when calculating χ^2 , the quality of the fit becomes better as χ^2 approaches 1. Figure 2d shows the fitting results for the four transient states as well as their reconstructed structures based on the optimized Au–Au distances.

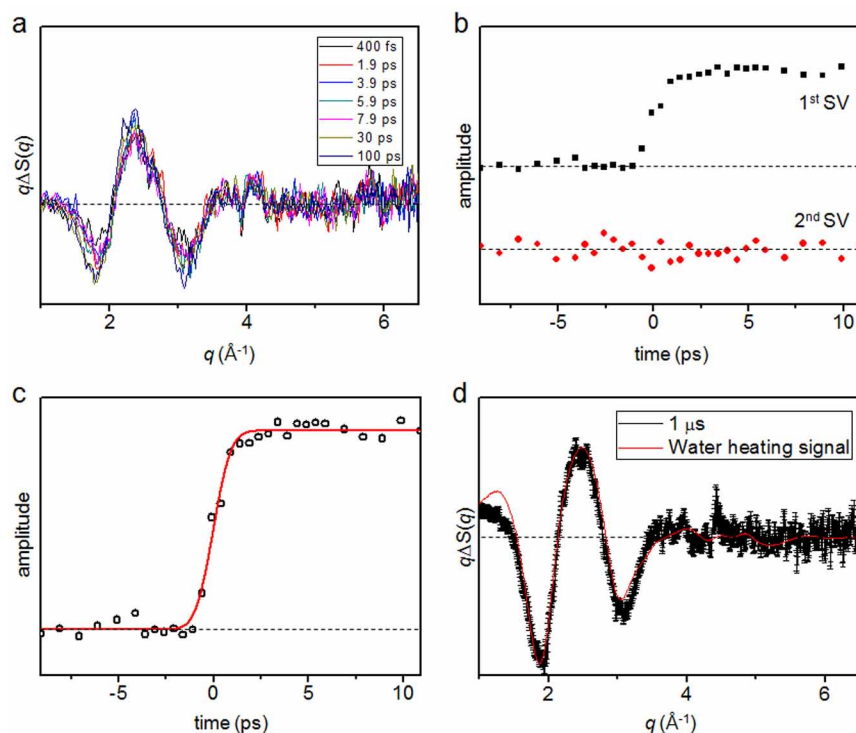
Determination of RDF of the S_0 state. As shown in Extended Data Fig. 4, time-independent, species-associated difference RDFs of the transient states were obtained by the SVD and kinetic analyses of the experimental, time-dependent, difference RDFs (Extended Data Fig. 3). These difference RDFs correspond to $r^2(S_{S_0}(r) - S_{S_0}(r))$, $r^2(S_{S_1}(r) - S_{S_0}(r))$, $r^2(S_{T_1}(r) - S_{S_0}(r))$ and $r^2(S_{\text{tetramer}}(r) - S_{S_0}(r))$ as indicated in Extended Data Fig. 4. From the structural fitting analysis, we were able to determine not only the RDFs of the transient states (S_1 , T_1 and the tetramer) but also the RDF of the S_0 state. We used a common S_0 structure when fitting the four species-associated difference RDFs. By optimizing the fit between the experimental and the theoretical difference RDFs for each transient species, we were able to obtain the theoretical RDF of the S_0 state in addition to the RDFs of the other states. To emphasize only the contributions of transient solute species associated with the bond formation, we added the RDF of the S_0 state to the experimental difference RDFs at all time delays and obtained the RDFs $r^2 S(r, t)$ shown in Extended Data Fig. 5.

- Inubushi, Y. *et al.* Determination of the pulse duration of an X-ray free electron laser using highly resolved single-shot spectra. *Phys. Rev. Lett.* **109**, 144801 (2012).
- Ishikawa, T. *et al.* A compact X-ray free-electron laser emitting in the sub-angstrom region. *Nature Photon.* **6**, 540–544 (2012).
- Rawashdeh-Omary, M. A., Omary, M. A. & Patterson, H. H. Oligomerization of $\text{Au}(\text{CN})_2^-$ and $\text{Ag}(\text{CN})_2^-$ ions in solution via ground-state aurophilic and argentophilic bonding. *J. Am. Chem. Soc.* **122**, 10371–10380 (2000).
- Kim, T. K., Lee, J. H., Wulff, M., Kong, Q. Y. & Ihee, H. Spatiotemporal kinetics in solution studied by time-resolved X-ray liquidography (solution scattering). *ChemPhysChem* **10**, 1958–1980 (2009).
- Ichihyanagi, K. *et al.* 100 ps time-resolved solution scattering utilizing a wide-bandwidth X-ray beam from multilayer optics. *J. Synchrotron Radiat.* **16**, 391–394 (2009).
- Haldrup, K. *et al.* Structural tracking of a bimolecular reaction in solution by time-resolved X-ray scattering. *Angew. Chem. Int. Ed.* **48**, 4180–4184 (2009).
- Jun, S. *et al.* Photochemistry of HgBr_2 in methanol investigated using time-resolved X-ray liquidography. *Phys. Chem. Chem. Phys.* **12**, 11536–11547 (2010).



Extended Data Figure 1 | Comparison of the TRXSS signals at SACLA and KEK and the TRXSS data in the entire time range. **a**, Comparison of the difference scattering curves at 100 ps time delay measured at SACLA (black) and KEK (red). The error bar at each data point indicates the standard error determined from 50 independent measurements. The two curves are nearly

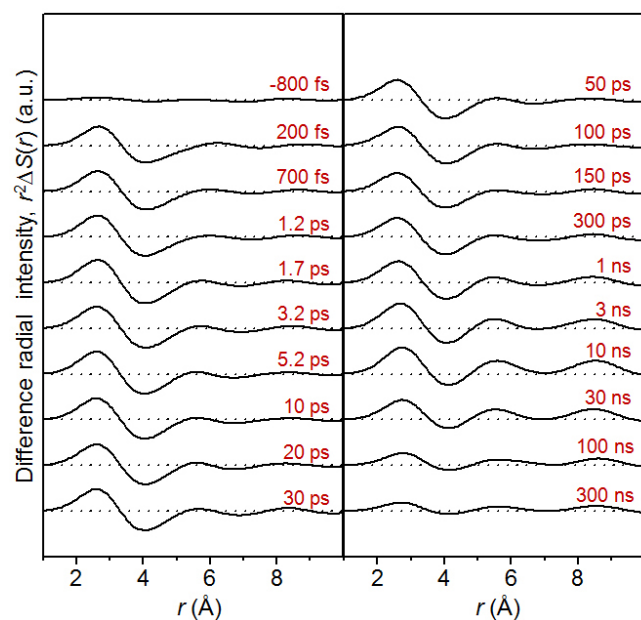
identical to each other within the experimental error, indicating that the difference scattering curves are highly reproducible and independent of the facility. **b**, Experimental difference scattering curves, $q\Delta S(q, t)$, in the entire time range from -800 fs to 1 μs.



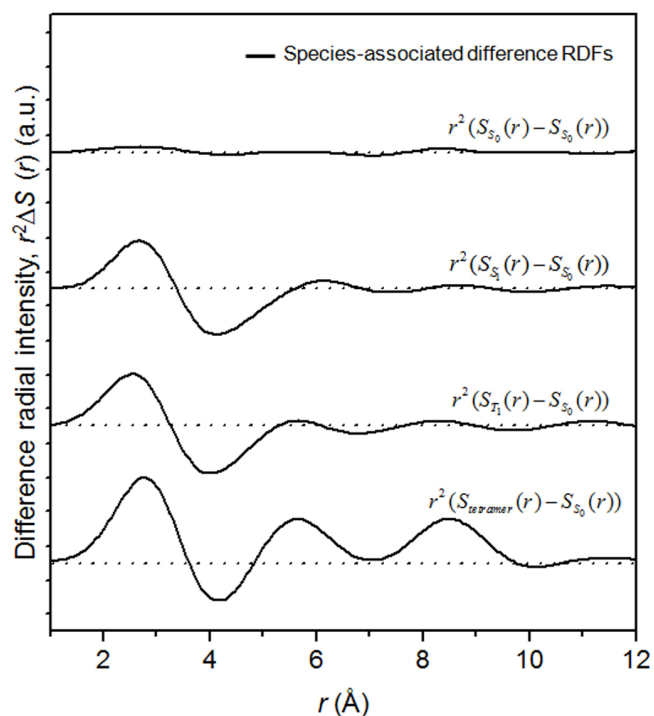
Extended Data Figure 2 | Solvent heating contribution to the TRXSS signal.

a, Experimental difference scattering curves, $q\Delta S(q)$, of FeCl_3 solution measured at several time delays (400 fs, 1.9 ps, 3.9 ps, 5.9 ps, 7.9 ps, 30 ps and 100 ps). **b**, SVD of the experimental difference scattering curves of FeCl_3 measured from -10 ps to 100 ps. The first two right singular vectors multiplied by singular values are shown. **c**, The first right singular vector (black circles) fitted by an error function (red curve). This result implies that only a single difference scattering curve accounts for solvent heating in the time range up to

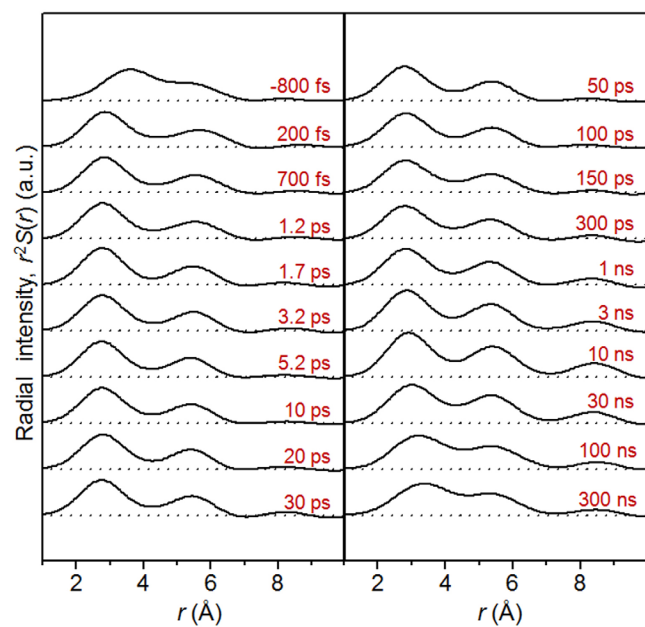
100 ps. **d**, Comparison of the difference scattering curve of the $[\text{Au}(\text{CN})_2]_3^-$ solution at $1 \mu\text{s}$ time delay (black) and the difference scattering curve for solvent heating (red). The error bar at each data point indicates the standard error determined from 50 independent measurements. At this time delay, the two curves are almost identical to each other within the experimental error, confirming that the difference scattering at late time delays are dominated by the solvent heating.



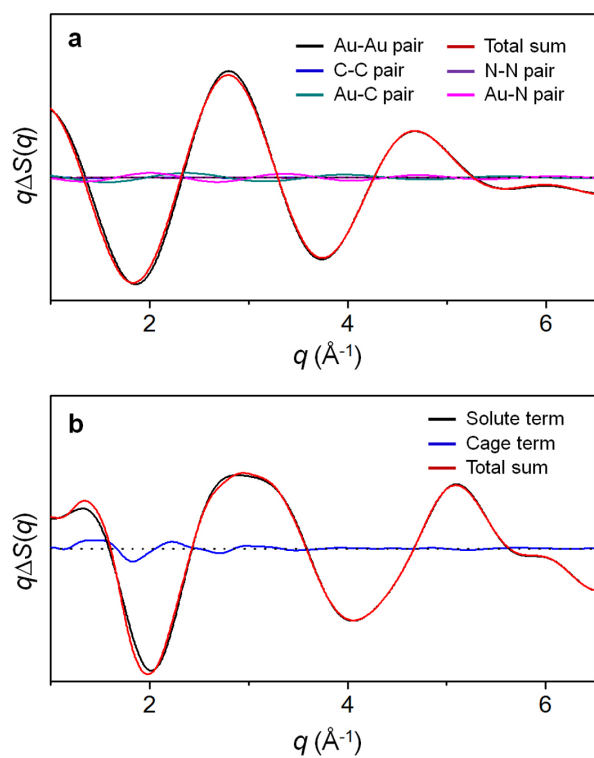
Extended Data Figure 3 | Difference RDFs in real space. Difference RDFs, $r^2\Delta S(r)$, obtained by Fourier sine transformation of $q\Delta S(q)$.



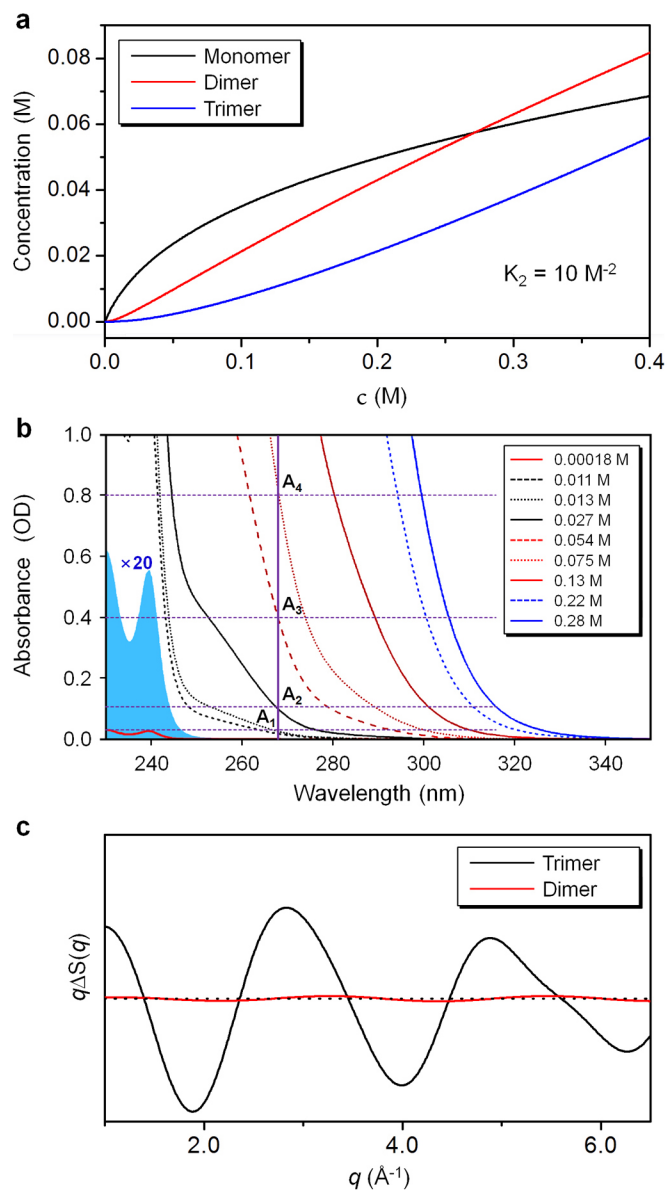
Extended Data Figure 4 | Species-associated difference RDFs of the transient states. The species-associated difference RDFs of the S_0 , S_1 , T_1 and tetramer states correspond to $r^2(S_{S_0}(r) - S_{S_0}(r))$, $r^2(S_{S_1}(r) - S_{S_0}(r))$, $r^2(S_{T_1}(r) - S_{S_0}(r))$ and $r^2(S_{tetramer}(r) - S_{S_0}(r))$, respectively. We used a common S_0 structure when fitting all four species-associated difference RDFs. By optimizing the fit between the theoretical and the experimental difference RDFs for each transient species via the structural fitting analysis, we were able to obtain the theoretical RDF of the S_0 state.



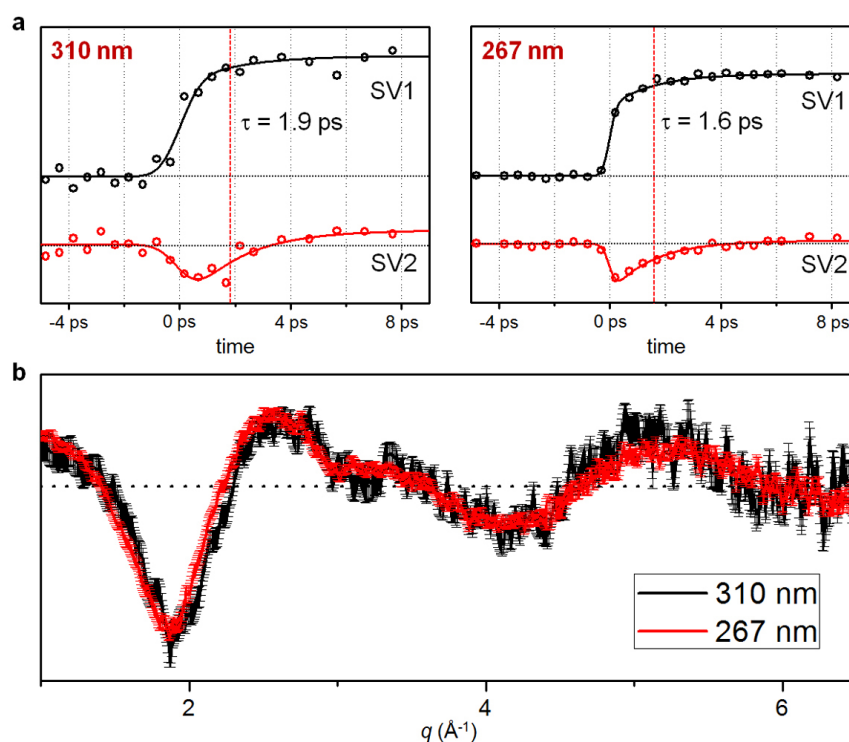
Extended Data Figure 5 | Radial distribution functions, $r^2S(r, t)$. The RDF of the S_0 state was added to the RDFs at all time delays to emphasize only the contributions of the transient solute species associated with the bond formation.



Extended Data Figure 6 | Comparison of the scattering from Au atoms and other contributions. **a**, Because the scattering intensities from C and N atoms are negligibly small, the total scattering pattern is almost the same as the scattering from Au atoms only. **b**, The contribution of the cage term is small and the total scattering pattern is therefore almost the same as the solute-only term.

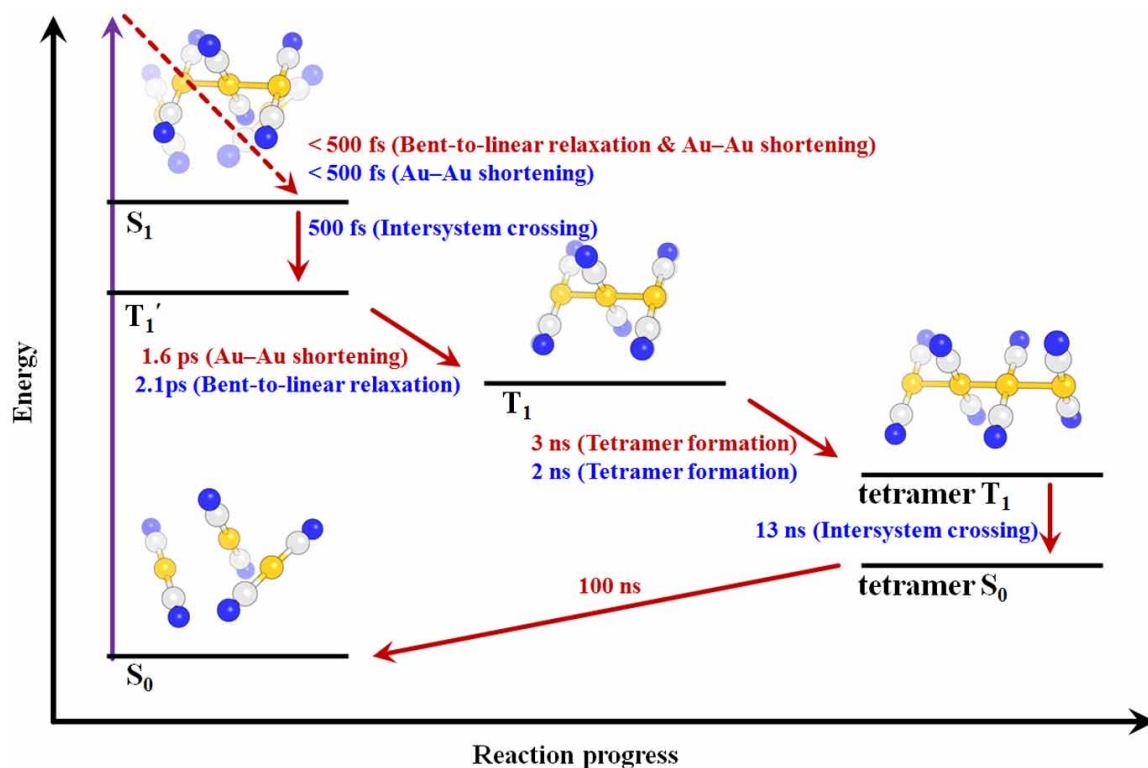


Extended Data Figure 7 | Contributions of trimer and dimer to X-ray scattering signal. **a**, Concentrations of the three species $[\text{Au}]$, $[\text{Au}_2]$ and $[\text{Au}_3]$, calculated as a function of c , which is the initial concentration of monomers of the gold complex. We assumed that K_2 is 10 M^{-2} in this case. **b**, Absorption spectra of aqueous solutions of $\text{K}[\text{Au}(\text{CN})_2]$ at various concentrations measured with a 0.5 mm path length cell. Four points (A_1 , A_2 , A_3 and A_4) that are used as inputs are indicated. **c**, Theoretical difference scattering curves for the trimer (black) and the dimer (red). Relative intensities of the two curves were estimated realistically based on the excitation probabilities and the equilibrium of the two species.



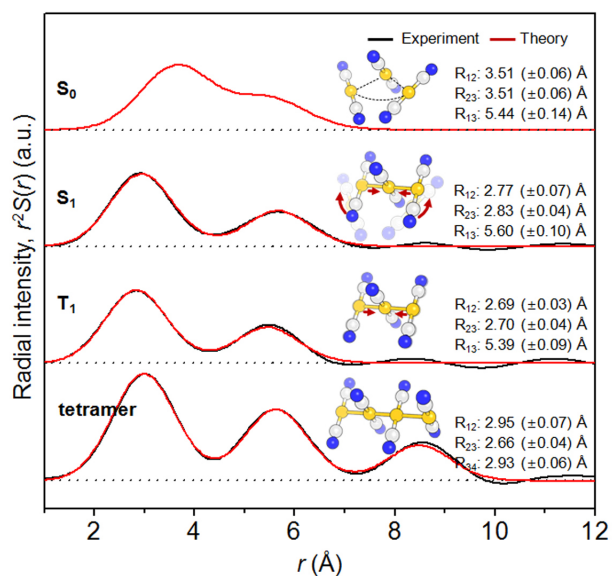
Extended Data Figure 8 | Comparison of the TRXSS data measured with excitations at two different wavelengths, 310 and 267 nm. **a**, Right singular vectors obtained from the SVD analysis for the 310 nm (left) and 267 nm (right) excitations. For both cases, the right singular vectors were fitted by only one kinetic component and their extracted time constants are almost identical to each other. **b**, Difference scattering curves at 100 ps time delay measured

with excitations at 310 nm (black) and 267 nm (red). The error bar at each data point indicates the standard error determined from 50 independent measurements. The two curves are identical to each other within the experimental error. This similarity between the kinetics and the shapes of the difference scattering curves indicates negligible contribution from dimer excitation for both 310 and 267 nm excitations.



Extended Data Figure 9 | Mechanism of photoinduced bond formation in [Au(CN)₂]³⁻. Results from our TRXSS data (red) and the previous transient absorption experiment (blue) are shown together. Our findings are in good agreement with the reaction mechanism proposed in the transient absorption study, except for the structural assignments of the early kinetics.

Considering that TRXSS is sensitive only to the processes accompanying structural change, the intersystem crossing processes on ~500 fs and 13 ns timescales, which were not observed in the TRXSS measurement, are likely to involve no structural change.



Extended Data Figure 10 | Species-associated RDFs of the four structures obtained from the SVD and principal-component analyses (black) and their fits (red) obtained by using the Debye–Waller factor and the constraint of the symmetric structure for the S_0 state. For each state, the structural parameters obtained from the fits and their standard errors determined from 50 independent measurements are shown together. It can be seen that the structural parameters of S_1 , T_1 and the tetramer obtained from the fits using the Debye–Waller factor and the symmetric constraint for S_0 are similar to the values given in Fig. 2d.

Extended Data Table 1 | Details of the data-collection parameters

Beamline	SACLA	SACLA	KEK
Date	2014. 4	2013. 9	2013. 5
X-ray energy (keV)	15.0	15.0	15.6
X-ray energy band width (%)	0.6	0.6	5
X-ray Size (mm ²)	0.2 × 0.2	0.09 × 0.08	0.2 × 0.2
X-ray fluence (mJ/mm ²)	1.30	1.87	0.017
X-ray pulse duration	< 100 fs	< 100 fs	100 ps
Spatial fluctuation (%)	10	10	~ 0
Repetition rate (Hz)	30	20	946
Laser wavelength (nm)	267	310	267
Laser pulse energy (μJ)	150	18	259
Laser size (vertical × horizontal, mm ²)	0.30 × 0.30	0.095 × 0.12	0.30 × 0.30
Laser fluence (mJ/mm ²)	2.12	2.05	3.67
Laser pulse duration	100 fs	100 fs	150 fs
Liquid jet thickness (mm)	0.1	0.1	0.3
Angle between X-ray and Laser (°)	10	10	10
Detector	Rayonix_LX255-HS	Rayonix_LX255-HS	MarCCD_165

Intensification and spatial homogenization of coastal upwelling under climate change

Daiwei Wang¹, Tarik C. Gouhier², Bruce A. Menge³ & Auroop R. Ganguly¹

The timing and strength of wind-driven coastal upwelling along the eastern margins of major ocean basins regulate the productivity of critical fisheries and marine ecosystems by bringing deep and nutrient-rich waters to the sunlit surface, where photosynthesis can occur^{1–3}. How coastal upwelling regimes might change in a warming climate is therefore a question of vital importance^{4,5}. Although enhanced land–ocean differential heating due to greenhouse warming has been proposed to intensify coastal upwelling by strengthening alongshore winds⁶, analyses of observations and previous climate models have provided little consensus on historical and projected trends in coastal upwelling^{7–13}. Here we show that there are strong and consistent changes in the timing, intensity and spatial heterogeneity of coastal upwelling in response to future warming in most Eastern Boundary Upwelling Systems (EBUSs). An ensemble of climate models shows that by the end of the twenty-first century the upwelling season will start earlier, end later and become more intense at high but not low latitudes. This projected increase in upwelling intensity and duration at high latitudes will result in a substantial reduction of the existing latitudinal variation in coastal upwelling. These patterns are consistent across three of the four EBUSs (Canary, Benguela and Humboldt, but not California). The lack of upwelling intensification and greater uncertainty associated with the California EBUS may reflect regional controls associated with the atmospheric response to climate change. Given the strong linkages between upwelling and marine ecosystems^{14,15}, the projected changes in the intensity, timing and spatial structure of coastal upwelling may influence the geographical distribution of marine biodiversity.

Coastal upwelling is a major oceanographic current that is prominent near the eastern boundaries of both the Atlantic and Pacific basins. In these EBUSs, coastal upwelling arises when equatorward winds along the eastern flanks of the subtropical highs transport surface waters offshore, causing them to be replaced by cold and nutrient-rich waters from depth via Ekman dynamics (Fig. 1a). The enhanced nutrient supply to the euphotic zone generated by these coastal upwelling currents sustains several productive fisheries and marine ecosystems around the globe: the California Current System (CalCS), off western North America; the Canary Current System (CanCS), off northwestern Africa and the Iberian Peninsula; the Humboldt Current System (HCS), off western South America; and the Benguela Current System (BCS), off southwestern Africa (Fig. 1a and Extended Data Fig. 1). Taken together, these four EBUSs cover less than 2% of the ocean surface but contribute 7% to global marine primary production and more than 20% to global fish catches¹⁶. The EBUSs span a wide range of latitudes and are therefore spatially heterogeneous environments. At higher latitudes, coastal upwelling is characterized by a marked seasonal cycle with the upwelling season beginning in spring and extending through summer and early autumn, whereas the winter season is dominated by downwelling. The length of the upwelling season increases progressively as latitude decreases, with upwelling becoming mostly a year-round phenomenon at tropical–subtropical latitudes (Fig. 1b–e). The timing, duration and intensity of coastal upwelling are known to have a critical role in the phenology of key marine

ecosystem processes such as the recruitment of rocky intertidal organisms, and changes in these upwelling characteristics have been shown to cause substantial disturbances to ecosystems at multiple trophic levels^{2,3,17}.

Climate change is expected to affect coastal upwelling and, thus, marine ecosystems in the EBUSs^{4,5}. Bakun proposed a mechanism whereby greenhouse warming would intensify the summertime alongshore winds and coastal upwelling by strengthening the land–sea thermal difference and surface pressure gradient in upwelling regions⁶. Subsequent analyses based on historical observations and palaeoclimate reconstructions have found evidence for increased upwelling-favourable winds in some parts of the EBUSs^{8,18,19} but not in others^{10,20}, leading to disagreements about coastal wind trends across different data sources^{9,10}. Climate model studies on projected changes to coastal upwelling have also yielded inconsistent results^{11–13}. Thus, there seems to be considerable debate regarding the impact of climate change on coastal upwelling⁴. A recent retrospective meta-analysis partially addressed this controversy by showing that coastal upwelling has intensified over the past 60 years²¹. Here we present a complementary prospective analysis using state-of-the-art climate models to understand how coastal upwelling will change under future greenhouse warming over the course of the twenty-first century. We use off-shore wind-driven Ekman transport as an index of coastal upwelling, and analyse historical and future simulations of 22 Earth system models developed for the Coupled Model Intercomparison Project phase 5²² (CMIP5) at multiple latitudes along the four EBUSs (Fig. 1a and Extended Data Fig. 1). These CMIP5 models reproduce the observed latitudinal variation in upwelling duration (Fig. 1b–e) and intensity (Fig. 1f–i) across all four EBUSs.

Considerable expansion of the upwelling season of several days per decade between 1950 and 2099 is evident at high latitudes in all four EBUSs (Fig. 2a, b). There are also noticeable differences between the Northern and Southern hemispheres and among the different EBUSs. The HCS and BCS in the Southern Hemisphere show larger and more consistent trends than do both the CalCS and the CanCS in the Northern Hemisphere. The trends in the Southern Hemisphere systems also increase with latitude and, in the case of the HCS, reach ~6 days per decade at the southernmost latitude, whereas the trends in the Northern Hemisphere systems vary non-monotonically with latitude by about 1 or 2 days per decade. Furthermore, the HCS and BCS exhibit trends of similar magnitude at common latitudes, whereas the CalCS and CanCS exhibit divergent trends. The positive trends are consistent and robust across models in the Southern Hemisphere systems, but inconsistent across models in the CalCS (Extended Data Fig. 3a–d). At lower latitudes, upwelling remains year-round between 1950 and 2099, and there is therefore no change in the duration of the upwelling season. Despite regional differences, the lengthening of the upwelling season at high latitudes in the EBUSs is a robust global response to greenhouse warming among the CMIP5 models. Both earlier onset and later termination of the upwelling season contribute comparably to the prolonged duration of seasonal upwelling in a warmer climate (Extended Data Fig. 2 and Extended Data Fig. 3e–l).

¹Sustainability and Data Sciences Laboratory, Department of Civil and Environmental Engineering, Northeastern University, Boston, Massachusetts 02115, USA. ²Department of Marine and Environmental Sciences, Marine Science Center, Northeastern University, Nahant, Massachusetts 01908, USA. ³Department of Integrative Biology, Oregon State University, 3029 Cordley Hall, Corvallis, Oregon 97331, USA.

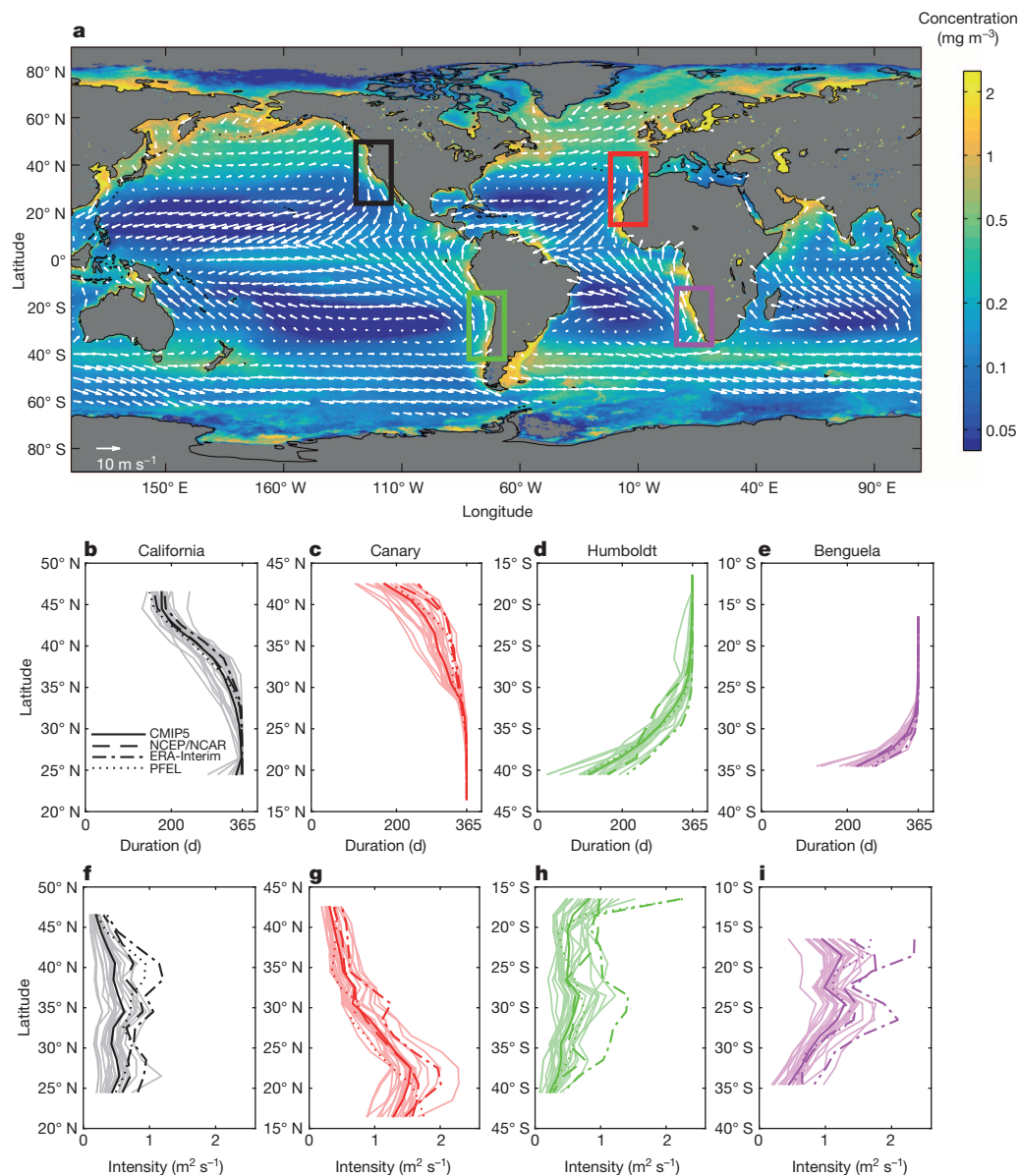


Figure 1 | Geographic locations of four EBUSs and latitudinal variations in coastal upwelling in each system. **a**, Aqua MODIS mean ocean chlorophyll *a* concentrations for 2002–2013 (colour scale); EBUS regions outlined with rectangular boxes and mean QuikSCAT ocean surface vector winds for 1999–2009 (white arrows). **b–i**, Mean durations of the upwelling season (**b–e**) and

upwelling intensity (**f–i**) for individual CMIP5 models (thin lines). Also shown are the multimodel mean (thick solid lines), the NCEP/NCAR reanalysis (thick dashed lines), the ERA-Interim reanalysis (thick dash-dot lines) and the PFEL upwelling index analysis (thick dotted lines) for 1981–2005 in each EBUS.

Global warming also has a strong and consistent effect on upwelling intensity, which we define as the average offshore Ekman transport over the upwelling season. Between 1950 and 2009, the CMIP5 models show a strengthening of upwelling at higher latitudes in all EBUSs except the CalCS, and weakening upwelling at lower latitudes in the CanCS (Fig. 2c, d). In the CalCS, the upwelling intensity exhibits modest weakening trends that are robust at three latitudes. In the HCS, strengthening trends are present at all latitudes but robust only at the three southernmost latitudes. These greenhouse-warming-induced trends in upwelling intensity are consistent and statistically robust across the climate models in the CanCS and BCS and, to a lesser degree, in the CalCS and HCS also (Extended Data Fig. 3m–p).

The increased duration and intensity of upwelling at higher latitudes and the lack of such trends at lower latitudes will reduce the latitudinal gradient of upwelling in the EBUSs. To demonstrate this effect, we computed the spatial standard deviation of each upwelling metric to quantify changes in the spatial heterogeneity of upwelling between 1950 and

2009. Higher and lower spatial standard deviation values indicate greater and, respectively, lower spatial heterogeneity. The spatial standard deviation of the duration of the upwelling season exhibits decreasing trends in all four EBUSs (Fig. 3a). The trend is stronger and more consistent across the models in the Southern Hemisphere systems than in the Northern Hemisphere systems (Extended Data Fig. 5a–d), consistent with the larger magnitude of the latitudinal trends in upwelling duration in the HCS and BCS (Fig. 2a, b). Comparable decreasing trends are evident in the spatial standard deviation of the upwelling season onset and termination dates (Extended Data Fig. 4). A similar reduction in spatial heterogeneity is also predicted for the upwelling intensity in all four EBUSs (Fig. 3b). The decreasing trends in the spatial standard deviation of upwelling intensity are most pronounced and consistent among climate models in the CanCS, followed by those in the BCS (Extended Data Fig. 5m–p), in line with the strong magnitude of the latitudinal trends in upwelling intensity (Fig. 2c, d). The reduction in the spatial heterogeneity of these upwelling characteristics is less prominent

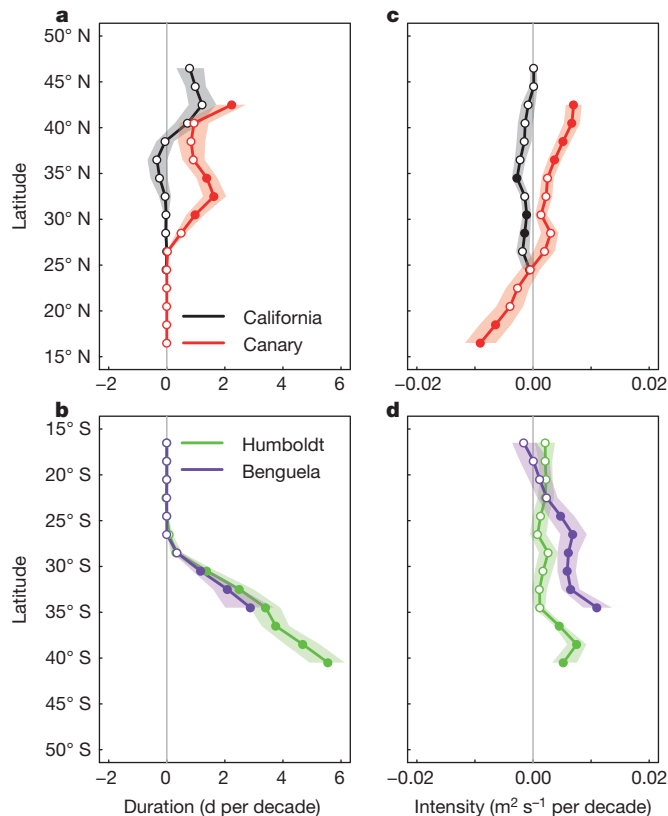


Figure 2 | Linear trends in upwelling duration and intensity. Multimodel means (solid lines) and 95% bootstrap confidence intervals (shading) of linear trends in upwelling duration (a, b) and intensity (c, d) for 1950–2099 for all four EBUSs. Filled circles represent trends that are robust across climate models (that is, at least 50% of the models show a statistically significant trend and at least 80% of those agree on the sign of the trend). The bootstrap confidence intervals are computed from 999 samples.

and robust in the CalCS and HCS. Nevertheless, the trend is statistically significant for half of the models and the multimodel mean (Extended Data Fig. 5).

The intensification of upwelling in both hemispheres over the course of the twenty-first century suggests that the underlying mechanism is related to global climate change. Bakun proposed that greenhouse warming would strengthen upwelling across the globe through differential land–sea surface heating because excessive summertime warming over land relative to the ocean intensifies the continental thermal lows adjacent to upwelling regions, thus increasing atmospheric pressure gradients and alongshore upwelling-favourable winds⁶. To test this hypothesis, we regressed the summertime upwelling intensity against the land–sea surface temperature difference at the high latitudes of the EBUSs between 1950 and 2099. Figure 4 shows that the increase in upwelling intensity is highly correlated with the increase in land–sea temperature difference in the CanCS, the HCS, the BCS and, to a lesser degree, the CalCS. This robust relationship between the land–sea temperature difference and upwelling intensity supports Bakun’s hypothesis, and suggests a link between greenhouse warming and the intensification of upwelling. The increase in coastal upwelling under climate change is also linked to changes in upwelling phenology because the onset and termination of the upwelling season correspond to the times of the year when the wind-driven coastal current changes from downwelling to upwelling, and vice versa. Hence, the projected year-round increase in upwelling-favourable winds causes the upwelling season to start earlier, end later and last longer (Extended Data Fig. 7).

Although the latitudinal trends in upwelling intensification are generally consistent across regions, certain patterns are region-specific and

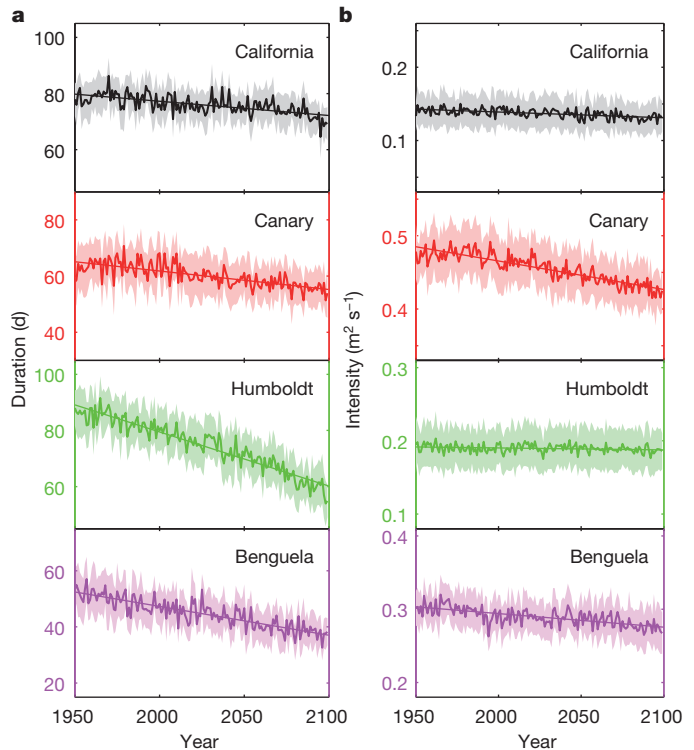


Figure 3 | Spatial standard deviations of upwelling duration and intensity. Multimodel means (thick lines) and 95% bootstrap confidence intervals (shading) of spatial standard deviation of upwelling duration (a) and intensity (b) for 1950–2099 for all four EBUSs. The thin straight lines indicate linear trends in the multimodel mean time series. The bootstrap confidence intervals are computed from 999 samples.

probably influenced by localized climate phenomena. For instance, the CalCS does not show intensification and the CanCS shows a robust weakening trend at the two lowest latitudes. Such differences among the EBUSs are probably the result of regional factors. Upwelling in the CalCS is strongly influenced by natural climate variability such as the El Niño/Southern Oscillation, Pacific Decadal Oscillation and North Pacific Gyre Oscillation^{23–26}, whose effects on upwelling may override those predicted by Bakun’s hypothesis. In the CanCS, an increase in the land–sea thermal difference is expected to strengthen the southwesterly monsoon circulation that drives downwelling-favourable winds in the subtropics, a mechanism that may explain the reduction in upwelling intensity at low latitudes²⁷. Additionally, the poleward shift of subtropical anticyclones also tends to weaken upwelling-favourable winds at low latitudes²⁸. Describing how such regional processes interact with global greenhouse forcing will be critical to further resolve the dynamics of upwelling in a warming climate.

The lengthening and strengthening of upwelling at high latitudes and the resulting reduction in its latitudinal heterogeneity may have profound ecological impacts. On local scales, the climate-mediated intensification of upwelling could promote the productivity of fisheries and marine ecosystems by bringing more nutrient-rich waters to the surface, where they can subsidize the base of the food web⁴. Alternatively, the increased supply of such nutrient-rich and oxygen-poor waters from depth could have adverse effects on marine life by allowing hypoxic conditions to develop over large swaths of the coastal ocean and causing mass die-offs²⁹. These contrasting effects of upwelling intensification both depend on the continued delivery of nutrient-rich waters to the surface via upwelling. However, increased solar heating due to greenhouse warming could enhance stratification, deepen the thermocline and thus prevent cool and nutrient-rich waters from being upwelled^{4,30}. Such a decoupling of upwelling from the supply of nutrient-rich waters would jeopardize the persistence of fisheries and the functioning of marine ecosystems.

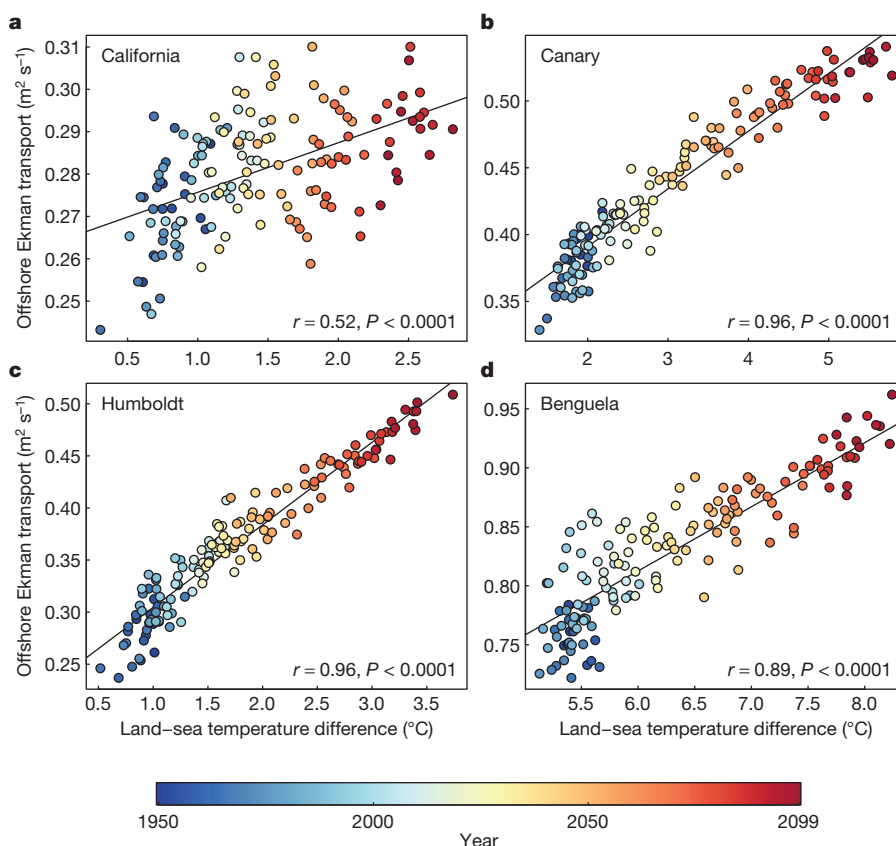


Figure 4 | Regression of summertime upwelling intensity against land-sea temperature difference. Scatterplots and generalized least-squares fits of multimodel mean summertime (May–September in the Northern Hemisphere and November–March in the Southern Hemisphere) offshore Ekman transport versus land-sea temperature difference for 1950–2099 at the three highest

latitudes in the CalCS (a), CanCS (b), HCS (c) and BCS (d). The correlations and corresponding P values between summertime upwelling intensity and land-sea temperature difference for individual models are tabulated in Extended Data Table 2.

On regional scales, the spatial structure of future upwelling trends is expected to alter the geographical distribution of species because both recruitment and biodiversity are strongly related to the existing latitudinal gradient in upwelling^{3,15}. Hence, by spatially homogenizing upwelling, climate change could remove an important environmental barrier associated with biodiversity, thereby increasing interspecific competition and species turnover in the coastal ocean. Overall, this suggests that the hotspots of climate-mediated change in upwelling identified here should be actively monitored to ensure the effective spatial management of productive fisheries and coastal ecosystems around the globe.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 30 July 2014; accepted 15 January 2015.

- Chavez, F. P. & Messié, M. A comparison of Eastern Boundary Upwelling Ecosystems. *Prog. Oceanogr.* **83**, 80–96 (2009).
- Barth, J. A. *et al.* Delayed upwelling alters nearshore coastal ocean ecosystems in the northern California current. *Proc. Natl Acad. Sci. USA* **104**, 3719–3724 (2007).
- Menge, B. A. & Menge, D. N. L. Dynamics of coastal meta-ecosystems: the intermittent upwelling hypothesis and a test in rocky intertidal regions. *Ecol. Monogr.* **83**, 283–310 (2013).
- Harley, C. D. G. *et al.* The impacts of climate change in coastal marine systems. *Ecol. Lett.* **9**, 228–241 (2006).
- Doney, S. C. *et al.* Climate change impacts on marine ecosystems. *Annu. Rev. Mar. Sci.* **4**, 11–37 (2012).
- Bakun, A. Global climate change and intensification of coastal ocean upwelling. *Science* **247**, 198–201 (1990).
- McGregor, H. V., Dima, M., Fischer, H. W. & Mülitz, S. Rapid 20th-century increase in coastal upwelling off northwest Africa. *Science* **315**, 637–639 (2007).
- García-Reyes, M. & Largier, J. Observations of increased wind-driven coastal upwelling off central California. *J. Geophys. Res.* **115**, C04011 (2010).
- Narayan, N., Paul, A., Mülitz, S. & Schulz, M. Trends in coastal upwelling intensity during the late 20th century. *Ocean Sci.* **6**, 815–823 (2010).
- Barton, E. D. D., Field, D. B. B. & Roy, C. Canary current upwelling: more or less? *Prog. Oceanogr.* **116**, 167–178 (2013).
- Mote, P. W. & Mantua, N. J. Coastal upwelling in a warmer future. *Geophys. Res. Lett.* **29**, 2138 (2002).
- Snyder, M. A., Sloan, L. C., Diffenbaugh, N. S. & Bell, J. L. Future climate change and upwelling in the California Current. *Geophys. Res. Lett.* **30**, 1823 (2003).
- Wang, M., Overland, J. E. & Bond, N. A. Climate projections for selected large marine ecosystems. *J. Mar. Syst.* **79**, 258–266 (2010).
- Blanchette, C. A. *et al.* Biogeographical patterns of rocky intertidal communities along the Pacific coast of North America. *J. Biogeogr.* **35**, 1593–1607 (2008).
- Fenberg, P. B., Menge, B. A., Raimondi, P. T. & Rivadeneira, M. M. Biogeographic structure of the northeastern Pacific rocky intertidal: the role of upwelling and dispersal to drive patterns. *Ecography* **38**, 83–95 (2015).
- Pauly, D. & Christensen, V. Primary production required to sustain global fisheries. *Nature* **374**, 255–257 (1995).
- Iles, A. C. *et al.* Climate-driven trends and ecological implications of event-scale upwelling in the California Current System. *Glob. Change Biol.* **18**, 783–796 (2012).
- Gutiérrez, D. *et al.* Coastal cooling and increased productivity in the main upwelling zone off Peru since the mid-twentieth century. *Geophys. Res. Lett.* **38**, L07603 (2011).
- Santos, F., Gomez-Gesteira, M., DeCastro, M. & Alvarez, I. Differences in coastal and oceanic SST trends due to the strengthening of coastal upwelling along the Benguela current system. *Cont. Shelf Res.* **34**, 79–86 (2012).
- Lemos, R. T. & Pires, H. O. The upwelling regime off the West Portuguese Coast, 1941–2000. *Int. J. Clim.* **24**, 511–524 (2004).
- Sydesman, W. J. *et al.* Climate change and wind intensification in coastal upwelling ecosystems. *Science* **345**, 77–80 (2014).
- Taylor, K. E., Stouffer, R. J. & Meehl, G. A. An overview of CMIP5 and the experiment design. *Bull. Am. Meteorol. Soc.* **93**, 485–498 (2012).
- Schwing, F. B., Murphree, T., deWitt, L. & Green, P. M. The evolution of oceanic and atmospheric anomalies in the northeast Pacific during the El Niño and La Niña events of 1995–2001. *Prog. Oceanogr.* **54**, 459–491 (2002).
- Di Lorenzo, E. *et al.* North Pacific Gyre Oscillation links ocean climate and ecosystem change. *Geophys. Res. Lett.* **35**, L08607 (2008).

25. Chenillat, F., Rivière, P., Capet, X., Di Lorenzo, E. & Blanke, B. North Pacific Gyre Oscillation modulates seasonal timing and ecosystem functioning in the California Current upwelling system. *Geophys. Res. Lett.* **39**, L01606 (2012).
26. Jacox, M. G., Moore, A. M., Edwards, C. A. & Fiechter, J. Spatially resolved upwelling in the California Current System and its connections to climate variability. *Geophys. Res. Lett.* **41**, 3189–3196 (2014).
27. Cropper, T. E., Hanna, E. & Bigg, G. R. Spatial and temporal seasonal trends in coastal upwelling off Northwest Africa, 1981–2012. *Deep-Sea Res. I* **86**, 94–111 (2014).
28. Belmadani, A., Echevin, V., Codron, F., Takahashi, K. & Junquas, C. What dynamics drive future wind scenarios for coastal upwelling off Peru and Chile? *Clim. Dyn.* **43**, 1893–1914 (2014).
29. Grantham, B. A. *et al.* Upwelling-driven nearshore hypoxia signals ecosystem and oceanographic changes in the northeast Pacific. *Nature* **429**, 749–754 (2004).
30. Roemmich, D., & McGowan, J. Climatic warming and the decline of zooplankton in the California Current. *Science* **267**, 1324–1326 (1995).

Acknowledgements This work was funded by grants from Northeastern University's Interdisciplinary Research Program and the US National Science Foundation's Expeditions in Computing program (award no. 1029711). We acknowledge the World Climate Research Program's Working Group on Coupled Modeling, which is responsible for CMIP5, and we thank the climate modelling groups for producing and making available their model output. For CMIP5 the US Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led the development of software infrastructure in partnership with the Global Organization for Earth System Science Portals.

Author Contributions T.C.G., D.W. and A.R.G. designed the study. D.W. and T.C.G. analysed the data. D.W. wrote the initial draft of the manuscript with substantial contributions from T.C.G. All authors discussed and interpreted the results and edited the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.W. (dw2116@gmail.com).

METHODS

CMIP5 model simulations. We analysed the output of the historical simulations and RCP8.5 simulations from 22 CMIP5 models including (1) ACCESS1.0, (2) ACCESS1.3, (3) BNU-ESM, (4) CanESM2, (5) CMCC-CESM, (6) CMCC-CMS, (7) CNRM-CM5, (8) CSIRO-Mk3.6.0, (9) GFDL-CM3, (10) GFDL-ESM2G, (11) GFDL-ESM2M, (12) HadGEM2-AO, (13) HadGEM2-CC, (14) IPSL-CM5A-LR, (15) IPSL-CM5A-MR, (16) IPSL-CM5B-LR, (17) MIROC5, (18) MIROC-ESM, (19) MIROC-ESM-CHEM, (20) MPI-ESM-LR, (21) MPI-ESM-MR and (22) MRI-CGCM3, where the number in parentheses in front of each model name is the model index used in Extended Data Figs 3, 5 and 8. The model selection is based on the output availability of daily near-surface winds. From the official CMIP5 data portal (<http://pcmdi9.llnl.gov/esgf-web-fe/>), we retrieved daily near-surface (10 m) wind speed (*sfWind*), zonal wind (*uas*), and meridional wind (*vas*) fields, as well as monthly surface temperature (*ts*) fields. The period analysed combines 1950–2005 from the historical simulations and 2006–2009 from the RCP8.5 simulations.

Offshore Ekman transport. We used daily offshore Ekman transport as an index of coastal upwelling, defined as the alongshore wind stress divided by the Coriolis parameter and seawater density. Because the CMIP5 multimodel ensemble does not provide output for daily or higher-frequency wind stress, we used the standard bulk aerodynamic formula to derive daily wind stress from CMIP5 daily near-surface wind speed, zonal wind and meridional wind with wind speed-dependent drag coefficient³¹. The near-surface winds were first projected onto the alongshore axis before being used to compute the alongshore wind stress via the bulk formula. In each EBUS, we compute daily offshore Ekman transport at an array of offshore locations evenly spaced by 2° in latitude and 1–2° (~100 km) from the coast (Extended Data Fig. 1 and Extended Data Table 1). Historical daily offshore Ekman transport values were also derived from the Pacific Fisheries Environmental Laboratory (PFEL) coastal upwelling index analysis³² and near-surface wind stresses from two atmospheric reanalysis products, namely the NCEP/NCAR reanalysis³³ and the ERA-Interim reanalysis³⁴. The PFEL coastal upwelling index has been extensively used to quantify coastal upwelling and its effects on fisheries and marine ecosystems at multiple trophic levels in the EBUSs^{3,15,17}.

Characterization of the phenology and strength of coastal upwelling. Following previous studies^{35,36}, we define the onset date, termination date, duration and intensity of the upwelling season as follows. We first computed the cumulative upwelling index (CUI) as the summation of daily offshore Ekman transport over a full year starting from boreal winter or austral winter, respectively 1 January and 1 July, at each latitude in the corresponding EBUSs. The onset of the upwelling season, also known as the spring transition, is the date on which the CUI reaches its annual minimum; likewise, the termination of the upwelling season, or the autumn transition, is the date on which the CUI reaches its annual maximum. The duration of the upwelling season is the total number of days between the onset date and termination date. The upwelling intensity is the average offshore Ekman transport over the upwelling season.

Relationship between the intensity and duration of coastal upwelling. Both upwelling intensity and duration are related to offshore Ekman transport, but in different ways. This is because upwelling intensity is defined as the ratio between cumulative Ekman transport during the upwelling season and the duration of the upwelling season, whereas upwelling duration is related to the time of the year when Ekman transport changes from being downwelling favourable to being upwelling favourable, or vice versa. Hence, for upwelling intensity to increase under climate change, the increase in cumulative Ekman transport over the upwelling season has to be greater than the increase in upwelling duration. If most of the increase in within-year Ekman transport occurs in the middle of the upwelling season, when Ekman transport is at its strongest, then there will be a tendency for upwelling intensity to increase but upwelling duration to remain the same. Conversely, if most of the increase in Ekman transport occurs at the onset or termination, when Ekman transport is at its weakest, then there will be a tendency for upwelling duration to increase but upwelling intensity to remain the same or even decrease.

Quantification of spatial heterogeneity of coastal upwelling. We quantified the spatial heterogeneity of the timing (onset and termination), duration and intensity of coastal upwelling in each EBUS and CMIP5 model using two different statistical measures. The first measure is the spatial standard deviation, which we defined for

each upwelling metric as $s = \sqrt{\sum_{i=1}^N (Y_i - \bar{Y})^2 / (N-1)}$, where Y_i is the value of the

upwelling metric at latitude i , \bar{Y} is the spatial mean of that metric averaged over all latitudes, and N is the number of latitudes in an EBUS. Large and small values of s are indicative of strong and, respectively, weak spatial heterogeneity, and an upward or downward trend in s represents increasing or, respectively, decreasing spatial heterogeneity. The second measure is the latitudinal slope coefficient, which we computed as the slope of a linear regression of each upwelling metric against the

absolute value of latitude in an EBUS: $\beta = \sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y}) / \sum_{i=1}^N (X_i - \bar{X})^2$. Here

X_i is the absolute value of the i th latitude and \bar{X} is the absolute value of the average latitude in an EBUS. The values of β are typically negative for the duration and termination dates of the upwelling season and upwelling intensity because these metrics tend to decrease with latitude within an EBUS. Conversely, the onset date tends to increase with latitude, resulting in positive β values. For all the upwelling metrics, trends in β towards and away from zero are indicative of decreasing and, respectively, increasing spatial heterogeneity. Compared with the spatial standard deviation, changes in the latitudinal slope coefficient reflect variations in the large-scale gradient in coastal upwelling. Both approaches indicate that the intensity, duration, onset and termination of upwelling are all becoming less spatially heterogeneous over time within each EBUS (Fig. 3 and Extended Data Figs 4 and 6).

Trend analysis. We determined the 1950–2009 trends and their significance (P values) by regressing the time series of each upwelling metric and its spatial standard deviation against time by the generalized least-squares method. This method was used instead of the more traditional least-squares method to account for autocorrelation in the time series. A trend value is regarded as statistically significant if the associated P value is less than 0.05. To ensure that our results were not sensitive to the statistical approach used to estimate trends, we also used the Mann–Whitney U test to determine whether the 50-year median value obtained for each upwelling metric differs between 1950–1999 (historical) and 2050–2099 (RCP 8.5). We used the Mann–Whitney U test because the upwelling phenology data are generally not normally distributed. The generalized least-squares and Mann–Whitney U analyses yielded similar results, confirming the robustness of our conclusions (Extended Data Figs 3 and 8).

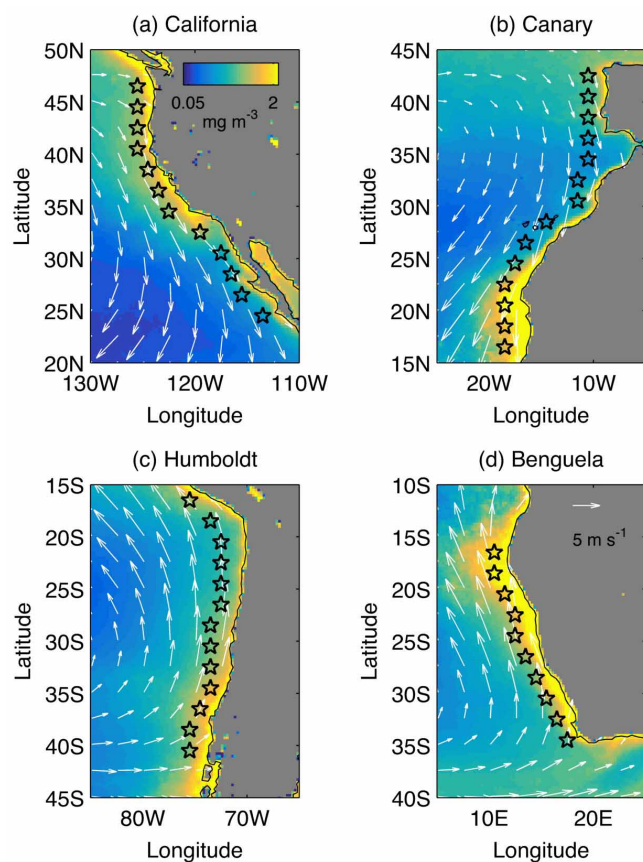
Robustness of upwelling trends across climate models. The projected changes in coastal upwelling reported in this study are based on an ensemble of 22 CMIP5 models. The robustness of these changes and our degree of confidence in them are commonly based on the level of model agreement³⁷. We quantified model agreement on upwelling trends using a method that assesses the degree of consensus based on both the statistical significance and the sign of the change³⁸. A change is interpreted as ‘robust’ when at least 50% of the models show a statistically significant trend and at least 80% of those agree on the sign of the trend. A trend is ‘not robust’ if (1) fewer than 50% of the models show statistically significant trend or (2) the majority of the models show a statistically significant trend but with low agreement on the sign of change. In this study, we chose to focus on the robust changes found in the majority of the EBUSs because they are most probably associated with global greenhouse warming.

Land–sea thermal difference and upwelling intensity. To determine the relationship between upwelling and land–sea thermal difference in the CMIP5 models, we computed the mean summertime (May–September in the Northern Hemisphere and November–March in the Southern Hemisphere) offshore Ekman transport over the three highest latitudes in each EBUS. We then determined the mean summertime land–sea surface temperature difference by computing the respective surface temperature differences between land and ocean averaged over a 10° × 5° region centred on the coastline across the CalCS (119° W–129° W, 42° N–47° N), CanCS (4° W–14° W, 38° N–43° N), HCS (68° W–78° W, 36° S–41° S) and BCS (13° E–23° E, 30° S–35° S) for each year. Finally, we regressed summertime offshore Ekman transport against land–sea surface temperature difference using the generalized least-squares method. These analyses were conducted for both the multimodel mean and each individual CMIP5 model (Fig. 4 and Extended Data Table 2).

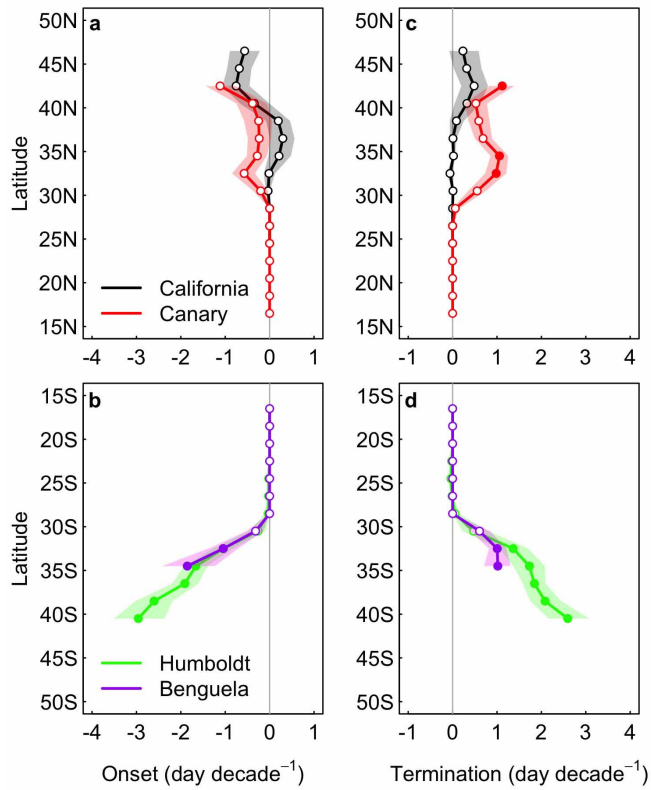
Coastal upwelling as a product of alongshore wind stress versus wind stress curl. Upwelling patterns in the EBUSs reflect both nearshore Ekman transport due to alongshore wind stress and offshore Ekman pumping due to wind stress curl^{39,40}. The typical spatial resolution of CMIP5 models is of the order of 100 km and is thus insufficient for resolving nearshore coastal upwelling patterns that occur within ~10 km of the coastline. Most existing studies, however, have used atmospheric observations or reanalysis products characterized by similar ~100 km resolutions to estimate coastal upwelling, and have successfully related these patterns to marine ecosystem functioning^{3,15,17}. The success of these low-resolution estimates of coastal upwelling may be attributable to their implicit integration of both nearshore Ekman transport and offshore Ekman pumping. Indeed, PFEL upwelling transport estimated from the 1°-resolution alongshore wind stress ~100 km offshore has a similar mean and variance to, and is highly correlated with, the total upwelling transport from a high-resolution (9 km) atmospheric model hindcast that comprises both nearshore Ekman transport due to alongshore wind stress and offshore curl-driven Ekman pumping⁴¹. This suggests that our CMIP5 upwelling estimates ~100 km offshore represent bulk upwelling transport that includes contributions from both nearshore Ekman transport and offshore Ekman pumping. Determining the relative contribution of these two processes to the overall patterns of upwelling will be an important next step for understanding how climate forcing influences both onshore and offshore oceanographic processes.

Data sources. Six-hourly US Navy Fleet Numerical Meteorology and Oceanography Center (FNMOC) Ekman transport data at 1° spatial resolution used to compute the PFEL upwelling index were retrieved from <http://coastwatch.pfeg.noaa.gov/erddap/griddap/erdlasFnTran6.html> and averaged into daily means. Daily zonal and meridional wind stress data in the NCEP/NCAR reanalysis product at 2.5° spatial resolution were retrieved from <http://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis.html>. Daily zonal and meridional wind stress data in the ERA-Interim reanalysis product at $\sim 0.7^\circ$ spatial resolution were retrieved from http://apps.ecmwf.int/datasets/data/interim_full_daily/. QuikSCAT climatological surface vector winds shown in Fig. 1 were retrieved from <http://cioss.coas.oregonstate.edu/scow/>. Aqua MODIS climatological chlorophyll *a* concentration data shown in Fig. 1 were retrieved from <http://oceanwatch.pifsc.noaa.gov/las/servlets/dataset?catitem=105>.

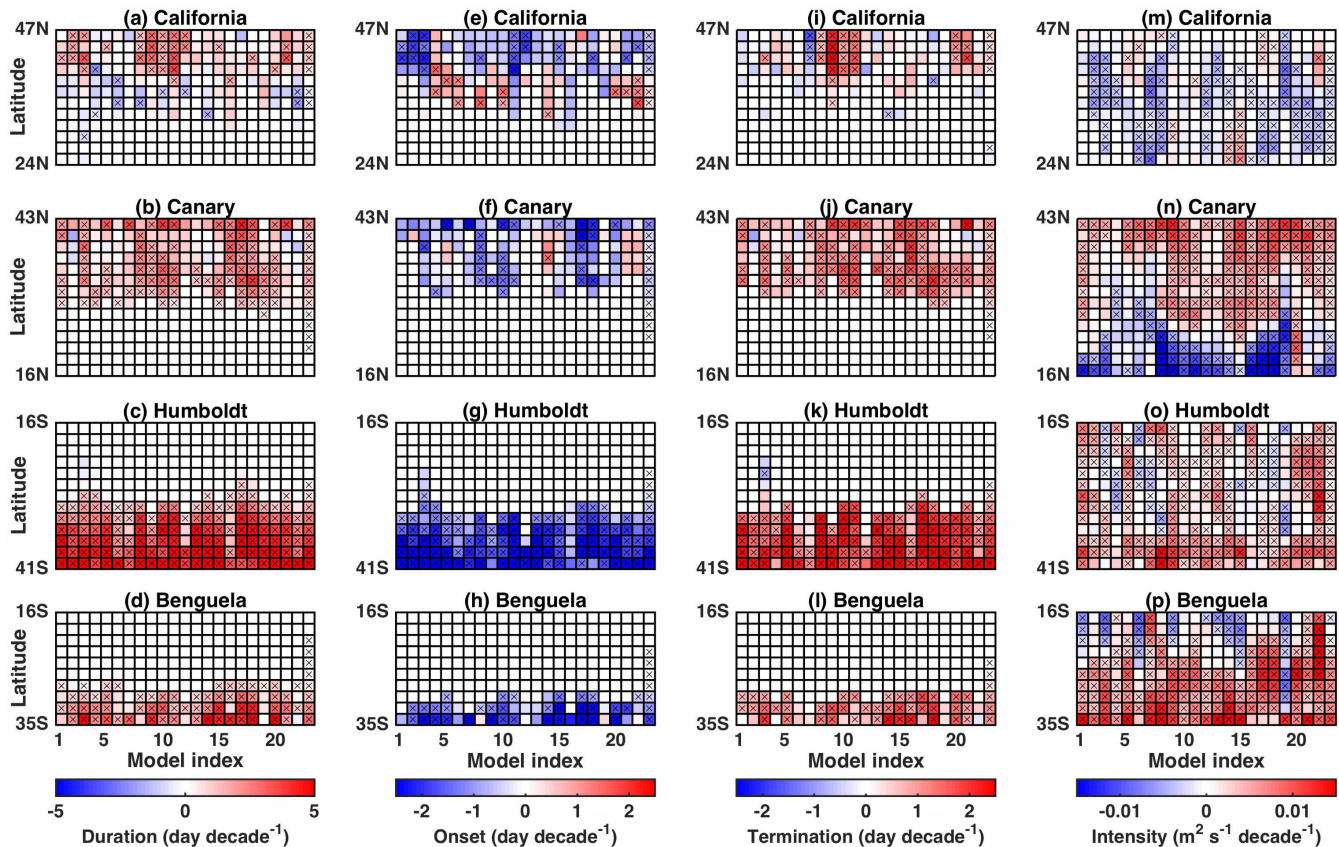
31. Trenberth, K. E., Large, W. G. & Olson, J. G. The mean annual cycle in global ocean wind stress. *J. Phys. Oceanogr.* **20**, 1742–1760 (1990).
32. Schwing, F. B., O'Farrell, M., Steger, J. M. & Baltz, K. *Coastal Upwelling Indices, West Coast of North America, 1946–1995*. NOAA Tech. Mem. NOAA-TM-NMFS-SWFC-231 (US Dept of Commerce, 1996).
33. Kalnay, E. *et al.* The NCEP/NCAR 40-year reanalysis project. *Bull. Am. Meteorol. Soc.* **77**, 437–471 (1996).
34. Dee, D. P. *et al.* The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Q. J. R. Meteorol. Soc.* **137**, 553–597 (2011).
35. Schwing, F. B. *et al.* Delayed coastal upwelling along the U.S. West Coast in 2005: a historical perspective. *Geophys. Res. Lett.* **33**, L22S01 (2006).
36. Bograd, S. J. *et al.* Phenology of coastal upwelling in the California Current. *Geophys. Res. Lett.* **36**, L01602 (2009).
37. Collins, M. *et al.* in *Climate Change 2013: The Physical Science Basis* (eds Stocker, T. F. *et al.*) 1029–1136 (Cambridge University Press).
38. Tebaldi, C., Arblaster, J. M. & Knutti, R. Mapping model agreement on future climate projections. *Geophys. Res. Lett.* **38**, L23701 (2011).
39. Rykaczewski, R. R. & Checkley, D. M. Influence of ocean winds on the pelagic ecosystem in upwelling regions. *Proc. Natl Acad. Sci. USA* **105**, 1965–1970 (2008).
40. Pickett, M. H. & Schwing, F. B. Evaluating upwelling estimates off the west coasts of North and South America. *Fish. Oceanogr.* **15**, 256–269 (2006).
41. Pickett, M. H. & Paduan, J. D. Ekman transport and pumping in the California Current based on the U.S. Navy's high-resolution atmospheric model (COAMPS). *J. Geophys. Res.* **108**, 3327 (2003).



Extended Data Figure 1 | Locations where daily offshore Ekman transport was computed along each EBUS. Also shown are Aqua MODIS mean ocean chlorophyll *a* concentrations for 2002–2013 (colour scale) and mean QuikSCAT ocean surface vector winds for 1999–2009 (white arrows) for the CalCS (a), CanCS (b), HCS (c) and BCS (d). The longitudes, latitudes and coast angles of all the locations (open stars) are given in Extended Data Table 1.

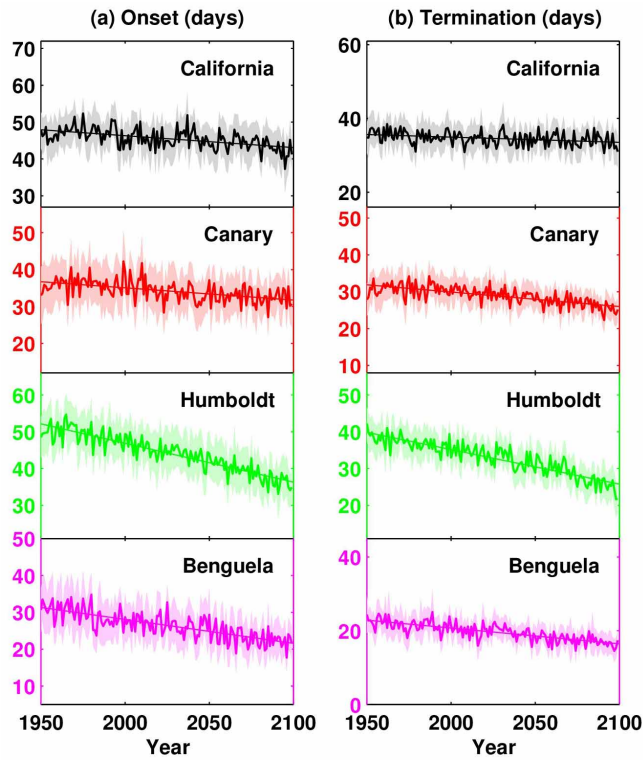


Extended Data Figure 2 | Linear trends in the timing of the upwelling season. Multimodel mean (solid lines) and 95% bootstrap confidence intervals (shading) of linear trends in the onset date (a, b) and termination date (c, d) of the upwelling season for 1950–2099 in all four EBUSs. Filled circles represent trends that are robust across climate models (that is, at least 50% of the models show a statistically significant trend and at least 80% of those agree on the sign of the trend). The bootstrap confidence intervals were computed from 999 samples.

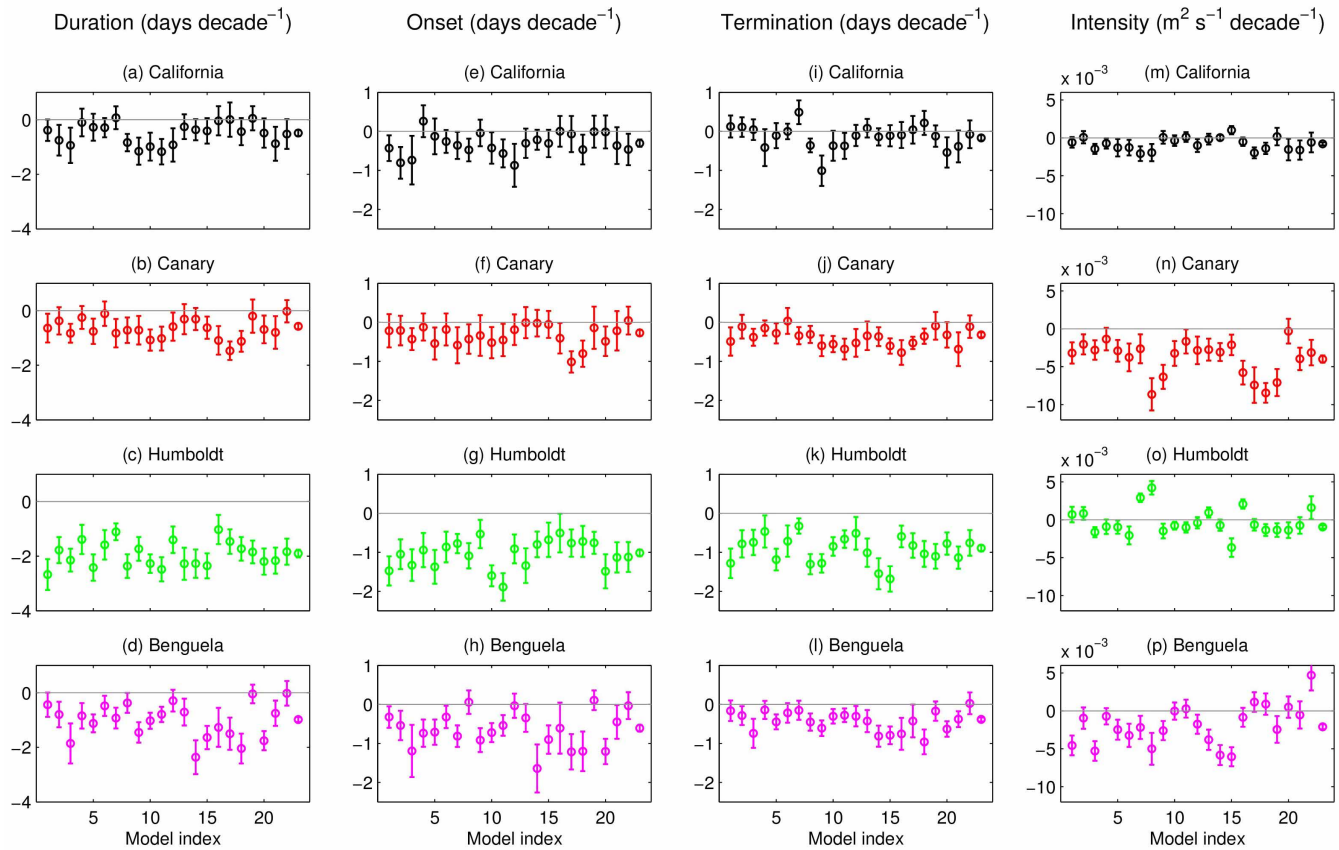


Extended Data Figure 3 | Linear trends in the upwelling metrics for the individual CMIP5 models. **a–d**, Generalized least-squares linear trends of upwelling duration for 1950–2099 in the CalCS (**a**), CanCS (**b**), HCS (**c**) and BCS (**d**). Red and blue respectively indicate positive and negative trend values. Crosses denote trend values that are statistically significant

(P value < 0.05). The first 22 columns are 22 CMIP5 models; the last column is the multimodel mean. **e–h**, Same as **a–d** but for the onset date of the upwelling season. **i–l**, Same as **a–d** but for the termination date of the upwelling season. **m–p**, Same as **a–d** but for upwelling intensity.

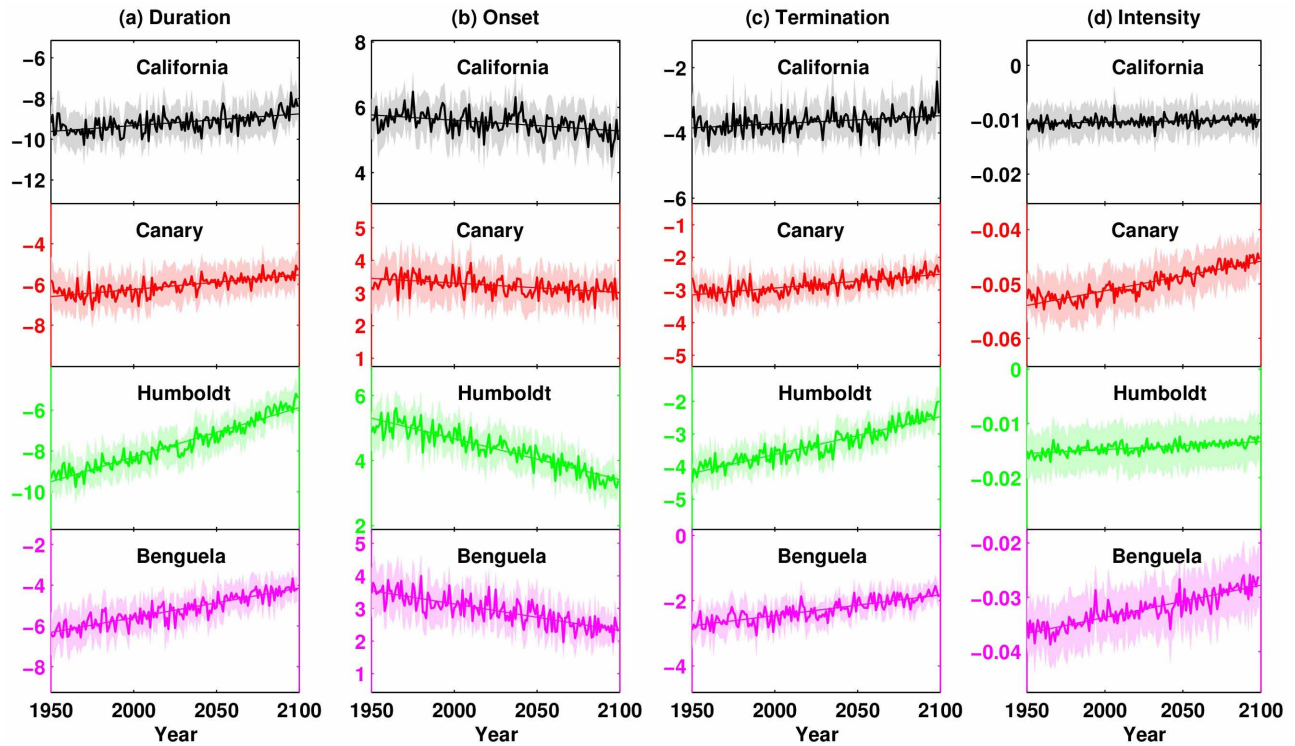


Extended Data Figure 4 | Spatial standard deviations of the timing of the upwelling season. Multimodel mean (thick lines) and 95% bootstrap confidence intervals (shading) of the spatial standard deviation of the onset date (a) and termination date (b) of the upwelling season for 1950–2099 in all four EBUSs. The thin straight lines indicate linear trends of the multimodel mean time series. The bootstrap confidence intervals are computed from 999 samples.



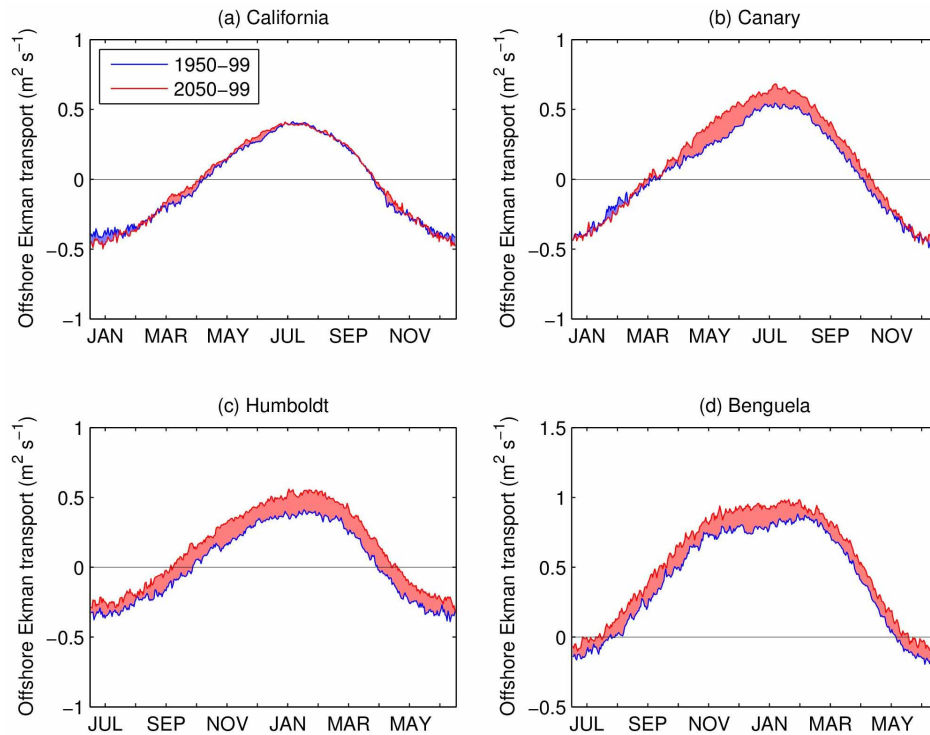
Extended Data Figure 5 | Trends in the spatial heterogeneity of coastal upwelling for individual CMIP5 models. **a–d**, Linear trends of the spatial standard deviation of the upwelling duration for 1950–2099 in the CalCS (a), CanCS (b), HCS (c) and BCS (d). Error bars indicate the 95% confidence

intervals. The first 22 bars are 22 CMIP5 models; the last bar is the multimodel mean. **e–h**, Same as **a–d** but for the onset date of the upwelling season. **i–l**, Same as **a–d** but for the termination date of the upwelling season. **m–p**, Same as **a–d** but for the upwelling intensity.



Extended Data Figure 6 | Latitudinal slope coefficients of upwelling metrics. Multimodel mean (thick lines) and 95% confidence intervals (shading) of the latitudinal slope coefficients of the duration (a; day per degree latitude), onset date (b; day per degree latitude) and termination date (c; day per

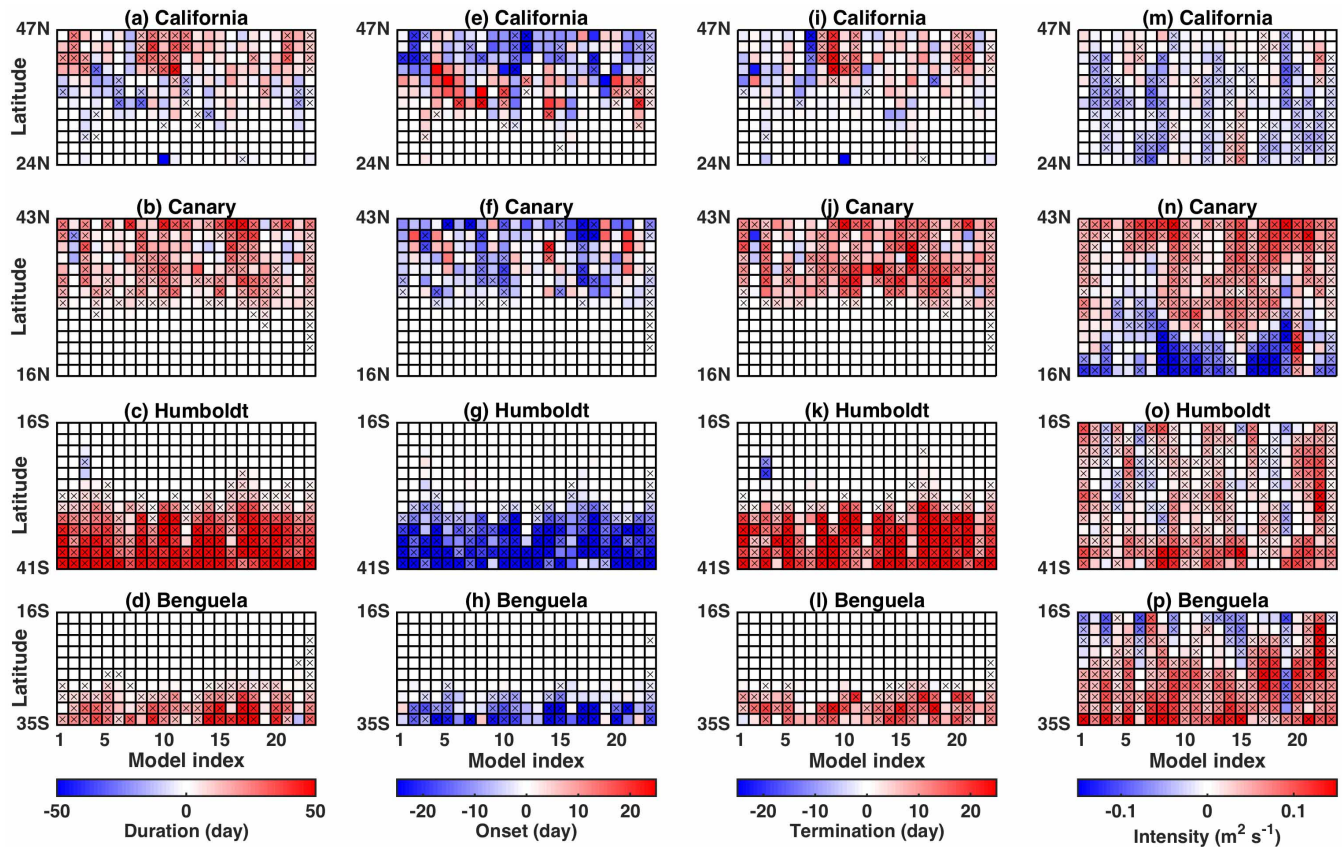
degree latitude) of the upwelling season, and of the upwelling intensity (d; $\text{m}^2 \text{s}^{-1}$ per degree latitude) for 1950–2099 in all four EBUSs. The thin straight lines indicate linear trends of the multimodel mean time series.



Extended Data Figure 7 | Intra-annual variation in upwelling trends.

Multimodel mean daily offshore Ekman transport for 1950–1999 (blue curve) and 2050–1099 (red curve) averaged over the three highest latitudes in the CalCS (a), CanCS (b), HCS (c) and BCS (d). Positive and negative temporal trends in the daily Ekman transport occur when the red curve is above and,

respectively, below the blue curve. These increases and decreases in upwelling transport are also highlighted by the red and, respectively, blue shading between the curves. The onset and termination of the upwelling season correspond to the times of the year when the daily Ekman transport first and, respectively, last reaches zero.



Extended Data Figure 8 | Median changes in upwelling metrics for the individual CMIP5 models. **a–d**, 50-year median change in the upwelling duration for 1950–1999 and 2050–2099 in the CalCS (**a**), CanCS (**b**), HCS (**c**) and BCS (**d**). Red and blue respectively indicate positive and negative trend values. Crosses denote changes that are statistically significant (P value < 0.05)

according to the Mann–Whitney U test. The first 22 columns are 22 CMIP5 models; the last column is the multimodel mean. **e–h**, Same as **a–d** but for the onset date of the upwelling season. **i–l**, Same as **a–d** but for the termination date of the upwelling season. **m–p**, Same as **a–d** but for the upwelling intensity.

Extended Data Table 1 | Latitudes, longitudes and coast angles of the locations representative of the EBUSs

California			Canary			Humboldt			Benguela		
Latitude	Longitude	Angle	Latitude	Longitude	Angle	Latitude	Longitude	Angle	Latitude	Longitude	Angle
24.5°N	113.5°W	122.2	16.5°N	18.5°W	80.8	16.5°S	75.5°W	132.1	16.5°S	10.5°E	98.0
26.5°N	115.5°W	112.9	18.5°N	18.5°W	78.8	18.5°S	73.5°W	124.7	18.5°S	10.5°E	111.0
28.5°N	116.5°W	117.8	20.5°N	18.5°W	71.5	20.5°S	72.5°W	104.2	20.5°S	11.5°E	108.9
30.5°N	117.5°W	126.0	22.5°N	18.5°W	59.4	22.5°S	72.5°W	96.0	22.5°S	12.5°E	102.8
32.5°N	119.5°W	134.0	24.5°N	17.5°W	54.2	24.5°S	72.5°W	92.7	24.5°S	12.5°E	105.7
34.5°N	122.5°W	116.1	26.5°N	16.5°W	47.9	26.5°S	72.5°W	86.5	26.5°S	13.5°E	111.1
36.5°N	123.5°W	116.2	28.5°N	14.5°W	69.5	28.5°S	73.5°W	88.6	28.5°S	14.5°E	114.0
38.5°N	124.5°W	112.9	30.5°N	11.5°W	76.7	30.5°S	73.5°W	86.9	30.5°S	15.5°E	108.9
40.5°N	125.5°W	100.1	32.5°N	11.5°W	68.0	32.5°S	73.5°W	82.8	32.5°S	16.5°E	106.3
42.5°N	125.5°W	88.5	34.5°N	10.5°W	79.2	34.5°S	73.5°W	76.5	34.5°S	17.5°E	133.1
44.5°N	125.5°W	88.4	36.5°N	10.5°W	88.8	36.5°S	74.5°W	83.4			
46.5°N	125.5°W	109.9	38.5°N	10.5°W	88.3	38.5°S	75.5°W	92.7			
			40.5°N	10.5°W	85.2	40.5°S	75.5°W	97.4			
			42.5°N	10.5°W	61.4						

The coast angles are defined relative to the zonal axis pointing east and are used to compute the alongshore component of surface wind stress.

Extended Data Table 2 | Correlations between upwelling intensity and land–sea temperature difference for individual CMIP5 models in all four EBUSs

	California		Canary		Humboldt		Benguela	
	<i>r</i>	<i>p</i> -value	<i>r</i>	<i>p</i> -value	<i>r</i>	<i>p</i> -value	<i>r</i>	<i>p</i> -value
ACCESS1.0	0.17	4.2E-08	0.24	2.9E-10	0.65	7.1E-34	0.38	1.4E-15
ACCESS1.3	0.20	7.2E-24	0.29	8.1E-13	0.67	3.5E-37	0.14	3.9E-07
BNU-ESM	0.37	1.2E-16	0.19	4.9E-09	0.36	7.0E-15	0.32	1.1E-11
CanESM2	0.09	7.9E-06	0.32	8.4E-14	0.52	1.9E-25	0.27	2.4E-11
CMCC-CESM	0.29	8.4E-13	0.46	4.8E-22	0.52	2.5E-25	0.38	3.6E-16
CMCC-CMS	0.47	3.3E-22	0.27	6.1E-10	0.39	9.0E-17	0.02	9.7E-02
CNRM-CM5	0.16	2.2E-07	0.21	2.0E-09	0.67	1.7E-35	0.48	1.8E-20
CSIRO-Mk3.6.0	0.26	2.0E-11	0.16	3.0E-07	0.84	2.8E-59	0.43	2.3E-20
GFDL-CM3	0.39	7.5E-18	0.54	3.6E-26	0.73	2.8E-42	0.03	5.4E-02
GFDL-ESM2G	0.29	1.6E-12	0.36	6.3E-14	0.41	6.4E-18	0.04	3.4E-02
GFDL-ESM2M	0.41	2.2E-22	0.31	1.6E-12	0.47	2.2E-20	0.04	3.7E-02
HadGEM2-AO	0.02	1.0E-01	0.05	6.7E-03	0.40	5.2E-18	0.14	1.9E-05
HadGEM2-CC	0.11	2.3E-06	0.24	1.5E-09	0.59	1.8E-30	0.41	4.0E-17
IPSL-CM5A-LR	0.33	8.9E-16	0.60	2.6E-29	0.76	2.6E-45	0.69	3.9E-34
IPSL-CM5A-MR	0.22	9.9E-15	0.62	1.1E-34	0.73	2.5E-45	0.58	5.2E-31
IPSL-CM5B-LR	0.35	2.0E-15	0.41	2.7E-19	0.47	2.9E-22	0.30	3.5E-12
MIROC5	0.32	1.2E-13	0.54	8.9E-25	0.32	3.5E-13	0.14	7.0E-06
MIROC-ESM	0.33	2.2E-15	0.57	2.4E-26	0.46	5.2E-24	0.16	8.7E-07
MIROC-ESM-CHEM	0.07	8.0E-17	0.49	1.1E-22	0.56	2.6E-31	0.08	3.7E-04
MPI-ESM-LR	0.21	2.2E-09	0.45	2.9E-20	0.57	2.2E-28	0.25	2.2E-10
MPI-ESM-MR	0.23	6.9E-10	0.47	1.5E-22	0.60	1.5E-28	0.27	3.2E-11
MRI-CGCM3	0.36	1.4E-16	0.28	4.9E-13	0.47	1.4E-22	0.23	1.6E-10
CMIP5 MME	0.42	1.1E-07	0.94	9.1E-58	0.95	1.7E-70	0.90	7.4E-37

The correlation coefficients *r* and corresponding *P* values are estimated by the generalized least-squares method. The results for the multimodel mean are given in the last row.

Seismic evidence of effects of water on melt transport in the Lau back-arc mantle

S. Shawn Wei¹, Douglas A. Wiens¹, Yang Zha², Terry Plank², Spahr C. Webb², Donna K. Blackman³, Robert A. Dunn⁴ & James A. Conder⁵

Processes of melt generation and transport beneath back-arc spreading centres are controlled by two endmember mechanisms: decompression melting similar to that at mid-ocean ridges and flux melting resembling that beneath arcs¹. The Lau Basin, with an abundance of spreading ridges at different distances from the subduction zone, provides an opportunity to distinguish the effects of these two different melting processes on magma production and crust formation. Here we present constraints on the three-dimensional distribution of partial melt inferred from seismic velocities obtained from Rayleigh wave tomography using land and ocean-bottom seismographs. Low seismic velocities beneath the Central Lau Spreading Centre and the northern Eastern Lau Spreading Centre extend deeper and westwards into the back-arc, suggesting that these spreading centres are fed by melting along upwelling zones from the west, and helping to explain geochemical differences with the Valu Fa Ridge to the south², which has no distinct deep low-seismic-velocity anomalies. A region of low S-wave velocity, interpreted as resulting from high melt content, is imaged in the mantle wedge beneath the Central Lau Spreading Centre and the northeastern Lau Basin, even where no active spreading centre currently exists. This low-seismic-velocity anomaly becomes weaker with distance southward along the Eastern Lau Spreading Centre and the Valu Fa Ridge, in contrast to the inferred increase in magmatic productivity¹. We propose that the anomaly variations result from changes in the efficiency of melt extraction, with the decrease in melt to the south correlating with increased fractional melting and higher water content in the magma. Water released from the slab may greatly reduce the melt viscosity³ or increase grain size⁴, or both, thereby facilitating melt transport.

Sea-floor spreading in the Lau back-arc system began about 4 Myr ago in the north of the basin⁵, propagated southwards and split the ancient arc into the Lau Ridge and the Tonga Ridge. This process formed the V-shaped Lau Basin, with segments of the spreading centres farther from the active Tofua volcanic arc in the north than in the south (Fig. 1a). This variation in distance from the arc correlates with systematic changes in the geological features of the spreading centres, with mid-ocean-ridge basalt (MORB)-like geochemical signatures in the northwest, and arc-like signatures and much higher inferred water content in the south^{1,6,7}. However, two questions remain unresolved. First, why is there an abrupt change in the geochemical signatures of subduction² and the structure of young ocean crust^{8,9} at about 20° 35' S along the Eastern Lau Spreading Centre (ELSC), despite the fact that the distance from the ridge to the arc changes gradually? Second, although the distance from the ELSC to the arc is intermediate between the respective distances from the Central Lau Spreading Centre (CLSC) and the Valu Fa Ridge (VFR) to the arc, why do the axial crustal properties of the ELSC imply the lowest magmatic activity⁶? These questions suggest complexities beyond the subduction-controlled melting process⁶. Previous studies rely on petrological and geochemical measurements of erupted basalts^{1,7}, sea-floor morphology⁵ or the structure of the crust^{8,9} to infer

characteristics of the melt production region in the mantle. In this study, we present new, high-resolution, three-dimensional seismic images of the Lau Basin, placing constraints on mantle melting variations in a region characterized by large gradients in mantle water content.

We used the two-plane-wave method of Rayleigh wave tomography^{10,11} and data from two networks of land and ocean-bottom seismographs (OBSs) (Fig. 1b and Extended Data Fig. 2) to image phase velocity at periods ranging from 19 to 88 s, and then determined the three-dimensional shear-wave velocity of the uppermost mantle. The derived azimuthally averaged velocity structure of vertically polarized S waves (SV waves) shows a wide low-velocity zone (LVZ) with a V shape in the shallow part of asthenosphere (Fig. 1b, c and Extended Data Fig. 1), dipping to the west away from the arc (Fig. 2). At shallow depths, for example 30 km, the LVZ occurs along the spreading centres and connects the CLSC to the Fonualei Rift and Spreading Centre and the Mangatolu Triple Junction (MTJ) to the north (Fig. 1c). At a depth of 50 km, the anomaly becomes stronger beneath the northeastern Lau Basin (NELB), but weaker beneath the southern ELSC and VFR (Fig. 1d). The Lau Ridge and the Fiji Plateau are characterized by a high-velocity anomaly in the uppermost mantle, implying that these relict island arcs are underlain by cold lithosphere to a depth of about 70 km. The high-velocity anomaly beneath the Tonga Ridge delineates the subducting Pacific slab¹² (Fig. 2).

The lowest mantle velocities are found at a depth of about 50 km along a band extending from the MTJ southwards to the northern tip of the ELSC, with a minimum SV-wave velocity of $3.5 \pm 0.15 \text{ km s}^{-1}$ (Fig. 1d), significantly lower than other well-studied upper-mantle LVZs^{13,14} (Methods and Extended Data Fig. 7). The velocity anomaly is considerably weaker and shallower to the south along the VFR (Fig. 3, D–D'). Because the southward decrease in anomaly magnitude correlates with a narrowing of the basin, the possibility that the decrease results from a lack of resolution for long-period Rayleigh waves must be considered. However, the anomaly trend is apparent even in mid-period phase velocity maps (for example those with a period of 37 s; Extended Data Fig. 3), with good resolution to the southern tip of the VFR (Extended Data Fig. 4). In addition, preliminary results from both independent analyses of shorter-period Rayleigh and Love waves using ambient-noise tomography¹⁵ (Extended Data Fig. 5) and body-wave attenuation¹⁶ provide evidence supporting a weaker low-velocity anomaly in the south (Methods).

Many factors, including temperature, composition and melt, influence seismic velocity^{17–22}. To investigate these effects, we estimated the SV-wave velocity structure of a 'wet' but melt-free mantle wedge beneath the Lau Basin based on numerical models²³, using experimental results fitted with an extended Burgers model²⁰ and corrections for the effects of water²⁴ and radial anisotropy¹⁴ (Methods and Extended Data Fig. 8). Although the predicted velocity structures (Extended Data Fig. 8) are similar to our observations (Fig. 2), the modelling is unable to explain the very low SV-wave velocities ($\leq 3.8 \text{ km s}^{-1}$) observed. The low

¹Department of Earth and Planetary Sciences, Washington University, St Louis, Missouri 63130, USA. ²Lamont-Doherty Earth Observatory, Columbia University, Palisades, New York 10964, USA. ³ Scripps Institution of Oceanography, University of California, San Diego, La Jolla, California 92093, USA. ⁴Department of Geology and Geophysics, University of Hawaii, Honolulu, Hawaii 96822, USA. ⁵Department of Geology, Southern Illinois University, Carbondale, Illinois 62901, USA.

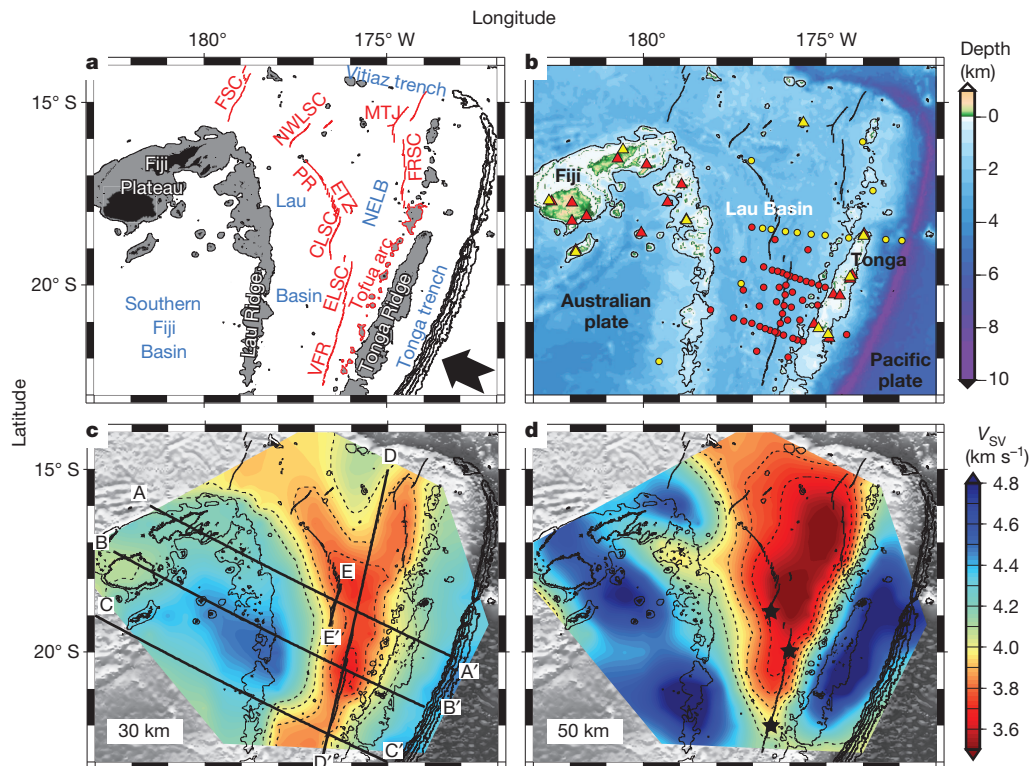


Figure 1 | Maps of the study region and mantle velocities. **a**, Tectonic map of the Lau Basin and adjacent areas with back-arc spreading centres (red lines). The Pacific plate subducts beneath the Tonga trench (delineated by water depth contours of 7, 8, 9 and 10 km) from the southeast (bold arrow). Land areas and water areas with depth shallower than 1 km are shaded in black and grey, respectively. Features with active magmatism have red labels. Volcanoes of the Tofua arc are outlined in red. ETZ, Extensional Transform Zone; FRSC, Fonualei Rift and Spreading Centre; FSC, Futuna Spreading Centre; NWLSC, Northwest Lau Spreading Centre. **b**, Seismic stations used in this study. A water depth (colour scale) of 1 km is shown by the contour. Red triangles represent

island-based stations operated from October 2009 to December 2010. Red dots are OBSs deployed from November 2009 to November 2010. Yellow dots and triangles indicate OBSs and island-based stations operated during September to December 1994, respectively (Extended Data Fig. 2). **c**, Azimuthally averaged SV-wave velocity (V_{SV} ; colour scale) at a depth of 30 km. Straight lines show the cross-sections in Figs 2 and 3. **d**, Azimuthally averaged SV-wave velocity at a depth of 50 km. Black stars are the nodes representing the CLSC, ELSC and VFR in the inset of Fig. 3 and Extended Data Fig. 3. In **c** and **b**, S-wave-velocity contours at 3.7, 3.8, 3.9, 4.0 and 4.1 km s^{-1} are shown by dashed lines. The water depth contours are the same in **b**.

velocities occur beneath the spreading centres at the depths expected for melt generation in the mantle (Fig. 3), but not in regions showing high water content, indicating that partial melt is the dominant factor. However, quantitative interpretation in terms of melt content is hampered by incomplete knowledge of the effects of melt on seismic velocity^{18,19}.

We suggest that the extent and intensity of the low-velocity anomaly provide constraints on the distribution and characteristics of the mantle melting process. The inclined LVZ (Fig. 2) shows a broad, asymmetric melting region originating at a depth of about 80 km, with the deepest part offset to the west, implying a passive decompression melting process governed by the mantle wedge flow pattern^{23,25}. Although it is difficult to quantitatively relate the low S-wave velocity observed beneath the CLSC to mantle porosity filled with melt (henceforth referred to as melt porosity), given the lowest S-wave velocity of about 3.9 km s^{-1} imaged beneath the East Pacific Rise¹³, the porosity here is higher than that beneath a fast-spreading mid-ocean ridge. This implies that the process of melt segregation at the CLSC is less efficient, consistent with the abnormally shallow depth of last melting equilibrium recorded in the lavas (about 35 km), because perfect segregation favours retention of high-pressure chemical signatures (Methods). The rapid spreading rate (about 90 mm yr^{-1} , that is, a high mantle matrix ascending rate) and high melt production due to high temperature both favour melt retention and high mantle porosities. Additionally, the higher spreading rate at the CLSC weakens the magmatic focusing at the ridge, hindering the extraction of melt²⁶.

The tomography results indicate that the MORB-like lavas erupting along the CLSC are derived from an upwelling zone (Fig. 2, A–A')

originating from the ambient mantle to the west of the Lau back-arc, well away from sources of water and fluid-mobile elements in the subducting slab. Although the northern ELSC is much closer to the slab than is the CLSC, there is still a connection of the LVZ to the west (Fig. 2, B–B'), suggesting that the source of melt may be dominated by ambient mantle near and west of the ELSC rather than by the subduction-influenced mantle. However, the VFR, characterized by high water content¹, shallow axial depths⁶ and anomalous major element compositions², lacks a sublithospheric melt zone to the west of the spreading axis (Fig. 2, C–C'). The horizontal component of corner flow in the mantle wedge beneath the VFR is probably slower than that beneath the CLSC²³, which may also lead to inefficient mantle supply from the west along the VFR. Therefore, although map-view results in Fig. 1 cannot resolve the details of this transition, owing to the low lateral resolution, the contrast between cross-sections B–B' and C–C' in Fig. 2 suggests that the sudden change in magma chemistry beginning at about 20° 35' S along the ELSC² represents the transition between spreading centres fed by decompression melting west of the axis and those dominated by flux melting near the Tonga slab (Fig. 2).

Interpreting the along-strike variation in seismic anomalies is complicated, because both temperature and water content change from north to south¹. The seismic anomaly variations are strongest at a depth of about 50 km, well beneath the lithosphere–asthenosphere boundary according to the half-space cooling model (Fig. 2, dotted curves), and so lithospheric cooling cannot be a cause of this north–south variation. Furthermore, the trend in the seismic velocity anomaly is opposite to the trend in the inferred source water content at the CLSC, ELSC and

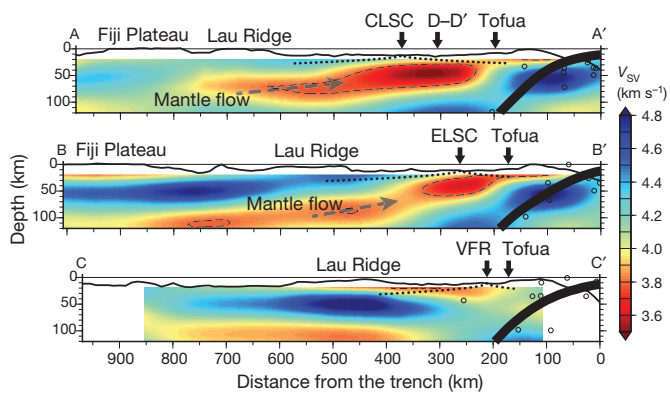


Figure 2 | Cross-sections A–A', B–B' and C–C' showing the azimuthally averaged SV-wave velocity. Owing to the low lateral resolution of surface waves at long periods, only structures shallower than 100 km depth are well resolved and, thus, interpreted. Local earthquakes (black circles; less than 100 km from each cross-section) located using the same data set delineate the surface of the subducting slab (thick curves; Slab 1.0 model²⁹). Dotted curves represent the bottom of the thermal lithosphere according to the half-space cooling model, demonstrating that most of the imaged velocity anomalies do not result from variations in conductive cooling. The bathymetry (solid curve along the top) is exaggerated vertically by a factor of five. The 3.8 km s^{-1} S-wave-velocity contour is shown by the dashed line. Beneath the CLSC and the ELSC, the decompression melting occurs in low-velocity regions interpreted as upwelling mantle from the west. In contrast, beneath the VFR in the south (C–C'), any connection between the ridge and the asthenosphere to the west is impeded by the lithosphere of the relict arc (wide blue anomaly beneath the Lau Ridge at 30–70 km depth), implying that the material supply of back-arc mantle from the west is much weaker and that the spreading centre samples only mantle in close proximity to the slab.

VFR, with areas of higher water content showing smaller and shallower seismic velocity anomalies (Fig. 3, inset). This contrast is surprising, given that the presence of water reduces the S-wave velocity in subsolidus olivine by enhancing attenuation²¹. We constrain possible thermal variations along-strike by using lava composition and the thermobarometer of ref. 27 to estimate the pressure (P)–temperature (T) conditions of melting (Fig. 3, Extended Data Fig. 9 and Methods). Calculated melting paths reveal that the mantle is hotter beneath the CLSC than beneath the VFR, possibly owing to less cooling by the slab. At 50 km depth, the temperatures beneath the CLSC and the VFR are about $1,400^\circ\text{C}$ and $1,350^\circ\text{C}$, respectively. This difference would cause a reduction in shear-wave velocity by only $<0.1 \text{ km s}^{-1}$ (Methods), much less than our observation of $>0.3 \text{ km s}^{-1}$.

Therefore, we interpret the along-strike variations in seismic anomalies in terms of changes in melt porosity. Our results suggest that the melt porosity is highest beneath the NELB and the CLSC, and from there decreases southwards to the VFR. In contrast, higher melt production is not expected for the NELB in numerical models²³, and as subaxial water content increases southwards, the onset of partial melt would be expected to deepen¹ and the extent of melting would increase towards the VFR, opposite to our inferences (Fig. 3). However, melt porosity and extraction are not expected to follow a simple relationship with melt productivity, but rather may be governed by other factors such as permeability and melt viscosity.

We propose that the variations in seismic velocities along the Lau back-arc spreading centres reflect differences in melt porosity due to changes in the efficiency of melt extraction, which determines the relation between melt content in the mantle and magmatic expression near the surface. The pattern of seismic anomalies suggests two major factors controlling the efficiency of melt transport. One factor is the

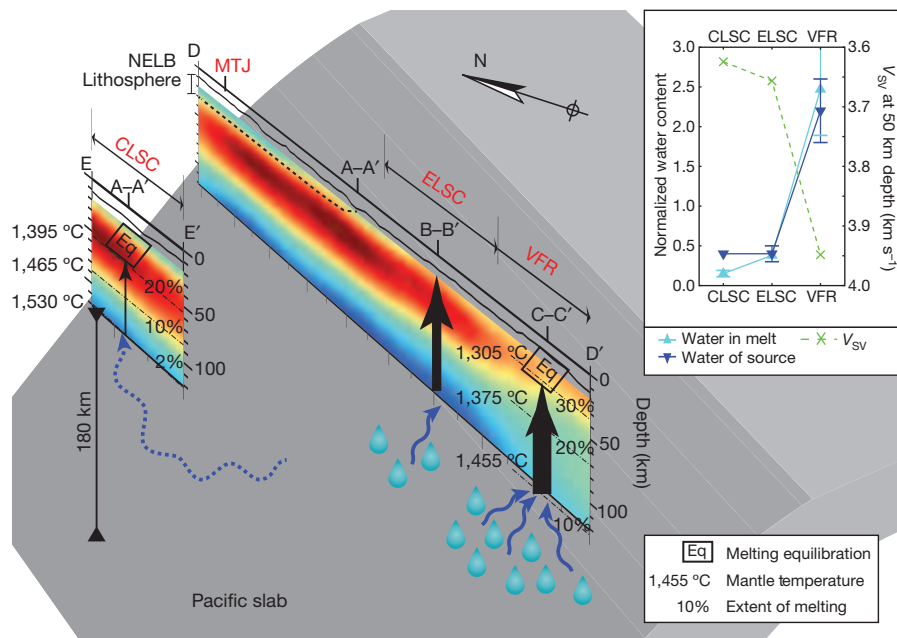


Figure 3 | Cross-sections D–D' and E–E' of azimuthally averaged SV-wave velocity with a schematic model showing the along-strike variations. The SV-wave velocity colour scale is the same as in Fig. 2. The dashed curve in D–D' represents a thin lithosphere with low permeability overlying the NELB. Mantle temperatures, the extent of melting and the depth of melt equilibration (boxed 'Eq') beneath the CLSC and the VFR are estimated from the thermobarometer of ref. 27, on the basis of the Si and Mg concentrations of primary melts in equilibrium with mantle olivine + orthopyroxene (Methods). The melting path beneath the ELSC is not available, owing to the limited data for water in primitive melts. Black arrows in the cross-sections represent upward melt transport, and blue wavy lines indicate water migration. In the north, where the CLSC is far from the slab, melts equilibrate at the depth of about 35 km and the

maximum extent of melting is about 25%. In contrast, relatively colder mantle with greater water content beneath the VFR generates melts that equilibrate about 10 km shallower than do those at the CLSC, and with a total extent of melting that is about 10% higher. As the distance between the spreading centre and the Tofua arc decreases, more water enters the melting region, leading to an enhancement in melt extraction (a wider arrow represents more efficient melt transport). Inset, trends of water concentration in the melt (Methods and Supplementary Table 1) and the mantle source¹ compared with our results for SV-wave velocity at a depth of 50 km. Error bars indicate the standard deviation among all samples (absence of error bars indicates there is only one measurement). The SV-wave velocity of each spreading centre is chosen as the velocity of the corresponding node in Fig. 1d.

existence of a nearby spreading centre as a focus for upward melt transport. The lowest seismic velocities occur beneath the NELB at substantial distance from active spreading centres, suggesting that the low permeability of the overlying lithosphere and the lack of an effective magma channel prohibit melt extraction. This implies that melt generated in the mantle wedge beneath this region is not efficiently extracted, but instead either slowly solidifies or migrates laterally large distances to one of the spreading centres.

The second major factor affecting melt porosity is the water content of the melt. Despite the high degrees of melting of the mantle at the VFR, the seismic anomalies are weaker throughout the melting region than they are beneath the CLSC, indicating lower melt porosity and greater melt extraction efficiency beneath the VFR, where melts are wetter. Compared with the CLSC, the water-rich VFR melts apparently segregate from the mantle more efficiently, similar to the profile beneath the Tofua volcanic arc. Both the seismic images and the P - T calculations suggest that melt rises efficiently, ponds and re-equilibrates at the base of the thermal boundary layer (20–25 km depth) (Fig. 3). Although the presence of water enhances melting¹, it also reduces the melt viscosity³ and facilitates grain growth⁴. If we assume the melt transport to be an equilibrium porous flow, it follows Darcy's law with $q \propto d^2 \phi^n / \mu$, where q is the melt flux, d is the grain size, ϕ is the porosity, μ is the melt viscosity and n is about 2.6 (ref. 28). Thus, a decrease in melt viscosity or an increase in grain size (or both) caused by higher water content would decrease melt porosity for a constant melt flux. In sum, our results and analysis indirectly imply that water greatly enhances melt mobility.

Ideally one could quantitatively relate the melt porosity inferred from seismology to melt extraction models of spreading centres. Unfortunately, there is no well-established relation linking seismic velocity and melt porosity. Many factors such as the grain size²⁰ and the topology of the partial melt within the matrix²² also affect seismic velocity, and other factors such as the behaviour of 'wet' melt under high pressure are poorly understood. Further experimental studies may provide the necessary constraints to relate seismic images directly to factors controlling melt production and transport.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 26 April; accepted 20 November 2014.

Published online 2 February 2015.

1. Kelley, K. A. *et al.* Mantle melting as a function of water content beneath back-arc basins. *J. Geophys. Res.* **111**, B09208 (2006).
2. Escrig, S., Bézous, A., Goldstein, S. L., Langmuir, C. H. & Michael, P. J. Mantle source variations beneath the Eastern Lau Spreading Center and the nature of subduction components in the Lau basin–Tonga arc system. *Geochem. Geophys. Geosyst.* **10** (2009).
3. Giordano, D., Russell, J. K. & Dingwell, D. B. Viscosity of magmatic liquids: a model. *Earth Planet. Sci. Lett.* **271**, 123–134 (2008).
4. Karato, S. Grain growth kinetics in olivine aggregates. *Tectonophysics* **168**, 255–273 (1989).
5. Taylor, B., Zellmer, K., Martinez, F. & Goodliffe, A. Sea-floor spreading in the Lau back-arc basin. *Earth Planet. Sci. Lett.* **144**, 35–40 (1996).
6. Martinez, F. & Taylor, B. Mantle wedge control on back-arc crustal accretion. *Nature* **416**, 417–420 (2002).
7. Pearce, J. A. *et al.* Geochemistry of Lau Basin volcanic rocks: influence of ridge segmentation and arc proximity. *Geol. Soc. Lond. Spec. Publ.* **81**, 53–75 (1994).
8. Dunn, R. A. & Martinez, F. Contrasting crustal production and rapid mantle transitions beneath back-arc ridges. *Nature* **469**, 198–202 (2011).
9. Arai, R. & Dunn, R. A. Seismological study of Lau back arc crust: mantle water, magmatic differentiation, and a compositionally zoned basin. *Earth Planet. Sci. Lett.* **390**, 304–317 (2014).

10. Forsyth, D. W. & Li, A. in *Seismic Earth: Array Analysis of Broadband Seismograms* (eds Levander, A. & Nolet, G.) 81–97 (American Geophysical Union, 2005).
11. Yang, Y. & Forsyth, D. W. Regional tomographic inversion of the amplitude and phase of Rayleigh waves with 2-D sensitivity kernels. *Geophys. J. Int.* **166**, 1148–1160 (2006).
12. Zhao, D. *et al.* Depth extent of the Lau back-arc spreading center and its relation to subduction processes. *Science* **278**, 254–257 (1997).
13. Harmon, N., Forsyth, D. W. & Weeraratne, D. S. Thickening of young Pacific lithosphere from high-resolution Rayleigh wave tomography: a test of the conductive cooling model. *Earth Planet. Sci. Lett.* **278**, 96–106 (2009).
14. Nishimura, C. E. & Forsyth, D. W. The anisotropic structure of the upper mantle in the Pacific. *Geophys. J. Int.* **96**, 203–229 (1989).
15. Zha, Y. *et al.* Seismological imaging of ridge–arc interaction beneath the Eastern Lau Spreading Center from OBS ambient noise tomography. *Earth Planet. Sci. Lett.* **408**, 194–206 (2014).
16. Wei, S. S. *et al.* in *AGU 2013 Fall Meeting*, abstr. D123B-07 (American Geophysical Union, 2013).
17. Hammond, W. C. & Humphreys, E. D. Upper mantle seismic wave velocity: Effects of realistic partial melt geometries. *J. Geophys. Res.* **105**, 10975–10986 (2000).
18. Faul, U. H., Fitz Gerald, J. D. & Jackson, I. Shear wave attenuation and dispersion in melt-bearing olivine polycrystals: 2. Microstructural interpretation and seismological implications. *J. Geophys. Res.* **109**, B06202 (2004).
19. McCarthy, C. & Takei, Y. Anelasticity and viscosity of partially molten rock analogue: toward seismic detection of small quantities of melt. *Geophys. Res. Lett.* **38**, L18306 (2011).
20. Jackson, I. & Faul, U. H. Grain-size-sensitive viscoelastic relaxation in olivine: Towards a robust laboratory-based model for seismological application. *Phys. Earth Planet. Inter.* **183**, 151–163 (2010).
21. Karato, S.-I. in *Inside the Subduction Factory* (ed. Eiler, J.) 135–152 (Wiley, 2004).
22. Takei, Y. Effect of pore geometry on V_P/V_S : from equilibrium geometry to crack. *J. Geophys. Res.* **107**, 2043 (2002).
23. Harmon, N. & Blackman, D. K. Effects of plate boundary geometry and kinematics on mantle melting beneath the back-arc spreading centers along the Lau Basin. *Earth Planet. Sci. Lett.* **298**, 334–346 (2010).
24. Karato, S.-I. On the origin of the asthenosphere. *Earth Planet. Sci. Lett.* **321–322**, 95–103 (2012).
25. Conder, J. A., Wiens, D. A. & Morris, J. On the decompression melting structure at volcanic arcs and back-arc spreading centers. *Geophys. Res. Lett.* **29**, 1727 (2002).
26. Kohlstedt, D. L. & Holtzman, B. K. Shearing melt out of the Earth: an experimentalist's perspective on the influence of deformation on melt extraction. *Annu. Rev. Earth Planet. Sci.* **37**, 561–593 (2009).
27. Lee, C.-T. A., Luffi, P., Plank, T., Dalton, H. & Leeman, W. P. Constraints on the depths and temperatures of basaltic magma generation on Earth and other terrestrial planets using new thermobarometers for mafic magmas. *Earth Planet. Sci. Lett.* **279**, 20–33 (2009).
28. Miller, K. J., Zhu, W.-L., Montési, L. G. J. & Gaetani, G. A. Experimental quantification of permeability of partially molten mantle rock. *Earth Planet. Sci. Lett.* **388**, 273–282 (2014).
29. Hayes, G. P., Wald, D. J. & Johnson, R. L. Slab1.0: a three-dimensional model of global subduction zone geometries. *J. Geophys. Res.* **117**, B01302 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank P. J. Shore, Y. J. Chen and the captains, crew and science parties of the RVs *Roger Revelle* and *Kilo Moana* for data collecting; D. W. Forsyth, Y. Yang, G. G. Euler, D. Heeszel, X. Sun and W. Shen for helping with data processing; N. Harmon, C. Rychert, P. Skemer and B. M. Mahan for discussions; and N. Hu for support. IRIS PASSCAL and OBSIP provided land-based seismic instrumentation and OBSs, respectively. This work was supported by the Ridge 2000 Program under NSF grants OCE-0426408 (D.A.W. and J.A.C.), EAR-0911137 (D.A.W.), OCE-0426369 (S.C.W.), OCE-0430463 (D.K.B.) and OCE-0426428 (R.A.D.).

Author Contributions S.S.W., advised by D.A.W., analysed the seismic data. T.P. downloaded and analysed the geochemical data. S.S.W. and D.A.W. took the lead in writing the manuscript, and all authors discussed the results and edited the manuscript.

Author Information Raw seismic data are available at the Data Management Center of the Incorporated Research Institutions for Seismology (<http://www.iris.edu/dms/nodes/dmc>), under network IDs YL, Z1 and XB. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.S.W. (songqiaowei@wustl.edu).

METHODS

Data processing and inversion. Most of the data used in this study were collected from 49 broadband OBSs deployed from November 2009 to November 2010 and 17 island-based seismic stations operated from October 2009 to December 2010. We additionally used data from 14 OBSs of the Lau Basin Ocean Bottom Seismograph Survey (LABATTS) and 9 island-based stations of the Southwest Pacific Seismic Experiment (SPASE) collected during September to December 1994¹² (Fig. 1b and Extended Data Fig. 2).

On the basis of the Preliminary Determination of Epicentres (PDE) catalogue, we selected seismograms of 357 earthquakes with surface-wave magnitudes (M_s) larger than 4.5 and epicentral distances between 30° and 150° (Extended Data Fig. 2, inset). The good azimuthal distribution of earthquakes guarantees a large number of ray-crossings north and south of the station array, improving resolution in the northern and southern Lau Basin (Extended Data Fig. 4). The raw seismogram of each event was cut from the origin time of the earthquake to 12,000 s after. Prior to the tomographic inversion, data were downsampled to 1 Hz and instrument responses were removed. For each period of interest, we used a narrow-bandpass filter (fourth-order Butterworth, zero-phase shift) centred at the frequency of interest to filter the seismograms. The filtered data were then windowed manually to isolate the fundamental mode of the Rayleigh wave at each of 20 periods in the range 19–118 s. Noise in seismograms at long periods (>50 s) due to ocean swell and associated water pressure variations, as well as tilt caused by local currents, was removed by correcting the vertical channel with horizontal and pressure channels^{30–32}.

We then used the two-plane-wave method¹⁰ with two-dimensional (2D) Fréchet kernels¹¹ to invert phase velocity with isotropic and anisotropic components with periods in the range 19–88 s. Periods longer than 88 s were not used because the wavelengths are too long compared with the size of our array to provide good resolution. Unlike the traditional Rayleigh wave tomography based on ray theory, this method considers scattering effects of Rayleigh waves outside the study region by simplifying the scattered incoming wave as the sum of two interfering plane waves¹⁰. Additionally, by using 2D Fréchet kernels based on the starting model, scattering and multipathing effects within the study region can be also approximately addressed¹¹. Both the calculation of the 2D Fréchet kernels and the nonlinear tomographic inversion require a good starting model of velocity. In the first step that determines the average phase velocity at each period for the entire study region, we chose as starting model the anisotropic model of the Pacific¹⁴ (henceforth the NF89 model) with age ranging from 0 to 4 Myr. In the second step, we divided the study region into four subregions according to tectonic settings: Lau Basin, Fiji Plateau/Lau Ridge, Tonga Ridge and background. Then we used the average phase velocity as the starting model to invert the 2D phase velocity map, with nodes spaced at 186.5 km, at each period. In the third step, we refined the grid of nodes to spacings of first 124.3 km and then 58.7 km, and used the previously inverted phase velocity as the *a priori* model.

Our results show strong azimuthal anisotropy in this region, consistent with previous studies³³. For instance, at a depth of 50 km, the fast direction is trench-parallel beneath the Lau Basin but convergence-parallel beneath the Lau Ridge and the Fiji Plateau. Although the amplitude of anisotropy at each period varies by up to 5% depending on the degree of regularization, the isotropic components of phase velocity show only very small changes (<0.03 km s⁻¹) owing to the good azimuthal coverage of the ray paths. Because the main purpose of this work is to image the 3D structure of seismic velocity and the inferred partial melt, only the isotropic phase velocities were analysed in the next steps.

Subsequently, we inverted the azimuthally isotropic phase velocity at each node to determine the azimuthally averaged SV-wave velocity using a linearized method³⁴. We initially tried to invert the SV-wave velocity with different uniform starting models based on NF89 models¹⁴ or on previous waveform inversions³⁵. Because this linearized inversion may depend on the starting model, we later divided the study region into six subregions: Lau Basin, Fiji Plateau/Lau Ridge, Tonga Ridge, North Fiji Basin, South Fiji Basin and Pacific plate. The starting model for each subregion was adopted from previous studies of seismic refraction³⁶, body-wave tomography³⁷ and NF89 models¹⁴. Each individual node was then sorted into one of these subregions, and the corresponding starting model was used in the inversion for the SV-wave velocity structure at that node. Compared with the inversion results from a uniform starting model, the results inverted from starting models of six subregions are almost identical for mantle structure deeper than 30 km but improve a little for shallower structure that agrees with geological settings better. We therefore chose the latter technique to obtain the final results of SV-wave velocity. Furthermore, to test the robustness of the extremely low velocity east of the CLSC, we applied the Monte Carlo algorithm to invert the node that has the lowest velocity. Tests show that the extremely low velocity east of the CLSC is robust, and SV-wave velocity inverted by the linearized method³⁴ is reasonable (see next section and Extended Data Fig. 7).

Resolution of phase-velocity inversion. Inversion of phase velocity involves an inevitable trade-off between spatial resolution and model resolution. Although a shorter smoothing length with a finer grid of nodes could lead to more small-scale information (higher spatial resolution), the over-parameterized problem will be poorly solved in the inversion, resulting in lower model resolution and failing to provide useful details¹⁰. We thus chose the inversion parameters mainly on the basis of the model covariance and checkerboard tests (Extended Data Fig. 4). For the largest node spacing of 186.5 km, we used a smaller *a priori* standard deviation (0.05 km s⁻¹) and a larger smoothing length (200 km), giving results more damped to the starting model. As we reduced the spacing of nodes with subsequent iterations, inversion parameters for shorter periods were changed accordingly³⁸ so that phase velocity had more variability (*a priori* standard deviation as 0.15 km s⁻¹ and smoothing length as 80 km for the finest grid of nodes with spacing 58.7 km). The lack of spatial resolution at long periods is an intrinsic problem for surface-wave tomography owing to the large wavelength of the waves and the consequent great width of the Fréchet kernel. For instance, the inferred subducting Pacific slab with dip angle gentler than shown by the Slab 1.0 model²⁹ (Fig. 2) is an artefact due to the low resolution as the longer-wavelength Rayleigh waves smear the horizontal structure more at larger depth. Because we were able to obtain useful resolution up to 88 s, only phase velocity at periods shorter than 90 s were used for the next step of S-wave-velocity inversion.

Ambient-noise tomography (ANT) uses ‘seismic noise’ to invert phase velocity at shorter periods and to constrain shallower structures than are resolved by the two-plane-wave tomography (TPWT). ANT results¹⁵ of phase velocity at the period of 18 s (Extended Data Fig. 5a) are consistent with phase velocity at 21 s obtained from TPWT, both showing weaker signal of LVZ in the south. Furthermore, ANT results of SV-wave velocity at the depth of 30 km (Extended Data Fig. 5b) show great agreement with TPWT (Fig. 1c) not only in pattern but also in absolute values. The smooth transition from shorter periods in ANT to longer periods in TPWT suggests our phase-velocity inversion is robust. Additionally, high attenuation anomalies revealed by independent body-wave analysis¹⁶ have a similar pattern to the LVZ in this study, supporting our results of a weaker low-velocity anomaly to the south.

Phase velocities at the CLSC are consistently lower than those observed at the Mariana back-arc³⁹ and the East Pacific Rise 12–18° S (ref. 13), whereas only short-period phase velocities at the VFR are lower (Extended Data Fig. 3). In addition, there are significant along-strike changes in the magnitude of this anomaly. To test our results for phase velocity, we applied the traditional two-station method for two sets of earthquake-station pairs (Extended Data Fig. 6). The first set contains one earthquake at the Chile trench recorded by two stations north of the ELSC, and the second set has one earthquake at the Mariana trench recorded by two stations near the VFR. Extended Data Figs 6b and 6c show the seismograms of the fundamental modes of Rayleigh waves filtered (fourth-order Butterworth, zero-phase shift) to 37 s. Given the differences in epicentral distance and phase delay time shown in Extended Data Fig. 6, the average phase velocity between N01W and N03W is about 3.59 km s⁻¹, and that between A12W and S01W is about 3.67 km s⁻¹, consistent with phase velocity inverted by the TPWT (Extended Data Fig. 3). These estimations of phase velocity are certainly approximate but lend evidence supporting our results for phase velocity.

Robustness of SV-wave velocity inversion. Linearized inversion³⁴ provides a fast way to invert S-wave velocity from phase velocity. However, the fixed thicknesses of model layers used in this method may lead to bias in the inverted results. Therefore we additionally applied a Monte Carlo algorithm to test the robustness of the extremely low velocity east of the CLSC (Extended Data Fig. 7). This method generates an ensemble of models using random perturbations to both velocity and layer thickness in the linearized inverted model (starting model). It then calculates the dispersion curve for each model, and compares it with the original dispersion curve from the phase-velocity inversion. A ‘good’ model is defined by two factors: (1) it should be as smooth as the linearly inverted model, and (2) its corresponding dispersion curve should have similar mis-fit compared with the starting model’s dispersion curve. The ‘best’ model is defined as the ‘good’ model that has smallest mis-fit. Numerical experiments with the inversion for node 364 (Extended Data Fig. 7) show that although the best model varies from inversion to inversion, the average of 500 good models is robust and almost identical to the linearly inverted model. The largest standard deviation of S-wave velocity over all depths is 0.13 km s⁻¹, much smaller than the range of perturbation set as 15%, that is, about 0.6 km s⁻¹. We thus conservatively estimate the uncertainty of the lowest velocity as 0.15 km s⁻¹. It is worthwhile to notice that the LVZ of our results is shallower than that from a previous waveform inversion study³⁵ (Extended Data Fig. 7, green curve), agreeing with the geological setting better as melting commences 60–70 km beneath the passive mid-ocean ridge⁴⁰. That is because the waveform inversion³⁵ averages over the whole back-arc and parts of the arc and the Fiji Plateau. Therefore, we conclude that SV-wave velocity inverted by the linearized method³⁴ is reasonable.

After trying various starting models and inversion parameters, an unexpected high-velocity zone (HVZ) deeper than about 100 km consistently appears in the result model, mainly owing to the abnormally steep gradient of the dispersion curve beyond 30 s. Beneath node 364, the subducting slab is presumably about 200 km deep, making it difficult to explain the HVZ geologically. However, given the fact that a segment of the slab 100 km deep is only 100 km east of the node, and that the wavelength for a 70 s Rayleigh wave, which is most sensitive to the depth of 100 km, is about 300 km, the HVZ can be explained as a smearing effect between the 'fast' slab and the 'slow' back-arc basin. As discussed in the previous section, the intrinsic problem of low resolution at long periods is serious in the back-arc basin, because the lateral variations are too dramatic to be accurately resolved by the long-wavelength surface waves. We thus limit our interpretation to the velocity structure shallower than 100 km, but have plotted structures between 100 and 120 km in the cross-sections for reference (Fig. 2).

For the same reason, the artificially high phase velocity at long periods, and the abnormally steep gradient of the dispersion curve, may also result in a slight bias towards low values in the S-wave velocity at shallow depths. Nevertheless, the fact that phase velocities at periods of 20–40 s are lower than those along the East Pacific Rise (Extended Data Fig. 3) guarantees that the shear velocities at depths of 30–60 km are consistently lower than the East Pacific Rise.

It is worthwhile to examine whether the observed structures at depths of less than 100 km are also artefacts due to this lateral smearing, especially the LVZ extending westwards that is parallel to the dominant direction of the wave paths (Fig. 2, A–A' and B–B'). Extended Data Fig. 4 shows that even at the period of 50 s, most sensitive to the depth of about 70 km, the lateral resolution is still good enough to resolve structures with a length scale of 200 km. In contrast, the results at 66 s show that even a larger length scale of 300 km cannot be resolved at periods of 66 s and longer, which are most sensitive to depths of 100 km and deeper. Additionally, preliminary results from independent analyses of body-wave attenuation¹⁶ also reveal a connection between the CLSC and the asthenosphere beneath the Fiji Plateau, which is expected from the numerical models²³.

Predicting shear-wave velocity. The extended Burgers model constrained by experimental data²⁰ provides a link relating temperature to seismic velocity, incorporating the important effects of anelasticity. We first used recent 2D numerical mantle wedge flow models²³ to predict S-wave-velocity structures beneath the CLSC, ELSC and VFR solely due to thermal variations. We corrected the predicted isotropic velocities to compare with our observed SV-wave velocities by assuming that the isotropic S-wave velocity is the average of V_{SV} and the velocity of the horizontally polarized S wave (V_{SH}), and by using the radial anisotropic parameters of NF89 0–4 Myr model¹⁴ (V_{SH}/V_{SV} varies from 1.010 to 1.016). Water content affects seismic velocity via anelastic behaviour, by changing the characteristic frequency of anelastic relaxation. We adopted the assumption that $\eta_{gbs} \propto Cw^r$ (ref. 24), where η_{gbs} is grain-boundary viscosity, Cw is water content and $r = 2$, and applied it to the 'dry' parameterizations²⁰ to estimate water effects. Because the potential temperature used in the numerical models is 1,450 °C (ref. 23), consistent with the value revealed at the CLSC but slightly higher than that at the ELSC and the VFR³⁵, and the exponent constant r lies at the upper limit of the assumption in ref. 24 ($r = 1$ – 2), the predicted SV-wave velocity should be underestimated. We neglected the effects of other compositional variations in seismic velocity in this study, because previous geochemical studies suggest that the Mg# of the mantle matrix varies by less than ± 1 as a result of melt depletion⁴¹ (Supplementary Table 1). These variations potentially lead to a change of $\pm 0.015 \text{ km s}^{-1}$ in S-wave velocity⁴², much smaller than the observed along-strike velocity change of $>0.3 \text{ km s}^{-1}$ (Fig. 3, inset).

Extended Data Fig. 8 shows the calculated SV-wave velocity under the considerations of temperature, water and radial anisotropy. We acknowledge that several parameters for the numerical models²³, the extended Burgers model²⁰ and water effects²⁴ are poorly constrained. It is thus difficult to evaluate the uncertainty of Extended Data Fig. 8. But these estimations at least provide a quantitative and visual way to assess the importance of each physical property. However, previous studies of geology, petrology and geochemistry all expected a large amount of melt beneath the Lau basin^{1,8}, where the low velocities are imaged. Therefore, we believe that it is reasonable to interpret the extremely low shear-wave velocities as partial melting.

Melting paths and along-strike thermal variations. To estimate the temperature of the melting region beneath the Lau spreading centres, we applied a thermobarometer based on the Si and Mg contents of primary liquids in equilibrium with olivine + orthopyroxene²⁷. We downloaded data from PetDB (<http://www.earthchem.org/petdb>)^{43–48}, selecting only submarine glasses that have been analysed for H₂O and that are primitive enough to be related to primary mantle melts by olivine crystallization only. The latter condition was met by selecting only those samples with MgO greater than that calculated for the point of plagioclase appearance on the cotectic (parameterized from the relationships in fig. 10 of ref. 49, as a function of H₂O, where $\text{MgO}(\text{plagioclase-in}) = (8.18 - 0.93) \times \text{H}_2\text{O}$). This resulted in only

four samples from the CLSC, only one from the ELSC and six from the VFR (Supplementary Table 1).

Even after selecting the most primitive basalts (Supplementary Table 1), chemical compositions still need to be corrected for olivine crystallization before being input into the mantle melt thermobarometer. To accomplish this, equilibrium olivine was added stepwise to the most primitive erupted compositions. We first calculated the composition of olivine in equilibrium with the erupted basaltic liquid using an Fe–Mg exchange coefficient between olivine and liquid of $K_D(\text{Fe}/\text{Mg})_{\text{ol/liq}} = 0.3$, next added 1% of that olivine to the melt compositions, and then calculated a new equilibrium olivine for the next 1% added, and so on. This procedure iteratively continued until the calculated melt compositions are in equilibrium with mantle olivine, here assumed to have the composition of Fo90 (90 mol% Mg/[Mg + Fe]). Because only Fe^{2+} participates in this exchange relationship, the $\text{Fe}^{3+}/\text{Fe}^{2+}$ of the melt must be known. We estimated $\text{Fe}^{3+}/\sum\text{Fe}$ from the concentration of H₂O in each sample, using the relationship in fig. 3 of ref. 50 for MORB and back-arc magmas; this yielded values between 14% and 18% $\text{Fe}^{3+}/\sum\text{Fe}$. The corrected primary (Fo90) melt compositions were then input into the thermobarometer of ref. 27, and the resulting P – T values are interpreted here to represent the last pressure and temperature of equilibration of melts in the mantle. Because the assumption of the $\text{Mg}/[\text{Mg} + \text{Fe}]$ of the mantle directly affects the T calculated, we consider Fo90 to be a minimum given the significant melt fractions (extents of melting) involved in this region, and so the calculated temperatures are minima for each sample.

Results in Extended Data Fig. 9 show that the CLSC–ELSC melts record temperatures about 50–75 °C higher than those beneath the VFR, and reflect slightly higher pressures of equilibration (1.1 versus 0.75 GPa, or depths of about 35 versus 22 km). All pressures, however, are very low, and consistent with the top of the melting region in both cases. Such a condition could exist either because melts remained in equilibrium with the solid matrix during decompression, or because melts were extracted efficiently but pooled at the base of the thermal boundary layer and re-equilibrated there. We prefer the former interpretation for the CLSC, owing to the very low seismic velocities observed in the melting region (30–100 km) there, and the latter for the VFR, owing to the higher observed seismic velocities and very shallow melt equilibration (comparable to the thermal boundary layer or the arc Moho, probably a rheological boundary where melts stall).

To trace the shallow equilibration conditions back to those of the full melting region, we calculated melting paths for average CLSC and VFR primary melts. First, the extent of melting (F) was calculated for the final melt, from its P , T and H₂O, using the cryoscopic approach of ref. 49 and values for dT/dF from ref. 51 (3.7 °C per percentage degree of melting, appropriate to 2 GPa; see fig. 1 in ref. 51), the depression of the basalt liquidus temperature as a function of water (the 3-oxygen mole curve at 1 GPa, from ref. 52), and the peridotite/melt partition coefficient for H₂O (0.007) from fig. 8 of ref. 53. The potential temperature of the mantle was then calculated by adding the heat of fusion (assuming $dT/dF = 6$ °C per percentage degree of melting) for that melt fraction, and by projecting to 0 GPa along a solid adiabat with slope of 10 °C GPa^{-1} . The H₂O concentration in the initial mantle is calculated from the H₂O concentration in the primary melt, the partition coefficient and the degree of melting. The melting path was then calculated from the potential temperature and the initial water content, using the same cryoscopic approach as above, taking into account the heat of fusion and solid adiabat.

The potential temperature (T_p) calculated in this way is about 1,500 °C for the CLSC and about 1,475 °C for the VFR. These T_p are at the high end of those estimated for mid-ocean ridges globally, and this has been a long-standing observation for the Lau Basin^{1,35,49}. The T_p calculated using the method here for northern Lau glasses (1,449 °C for average MT) and 1,463 °C for the Fonualei Rift and Spreading Centre (FRSC)) are nearly identical to those calculated using an independent method based on $\text{Na}_{(\text{Fo90})}$ and $\text{Fe}_{(\text{Fo90})}$ ($T_p = 1,449 \pm 23$ °C; ref. 35). The full melting paths are shown in Extended Data Fig. 9, with the CLSC mantle melting path remaining about 50–75 °C hotter than the VFR mantle path throughout. The somewhat steeper trajectory for the VFR melting path reflects its higher water content, which leads to greater overall extents of melting despite the lower initial temperature. The overall degrees of melting are very high ($>30\%$ for the VFR glasses), but this is consistent with the eruption of boninite liquids at the VFR (for example $\text{SiO}_2 > 53\%$, $\text{TiO}_2 < 0.5\%$, $\text{MgO} > 8\%$), which are generally taken to reflect equilibration with refractory mantle that has melting past cpx-out⁵⁴.

At the depth of 50 km, the temperatures beneath the CLSC and the VFR are about 1,400 °C and 1,350 °C, respectively. It is important to examine the implication of this thermal difference of 50 °C to our seismic observations. According to the extended Burgers model²⁰, an increase of 50 °C at a depth of 50 km will lead to a decrease in shear-wave velocity of $<0.1 \text{ km s}^{-1}$, much less than our observation of $>0.3 \text{ km s}^{-1}$. Additionally, experimental data for melt-bearing samples¹⁸ show that attenuation becomes more dependent on temperature when melt is present; thus, an increase in temperature may cause a larger decrease in shear-wave velocity

than the melt-free model predicts²⁰. Following ref. 55, we estimated shear-wave velocity as a function of temperature T :

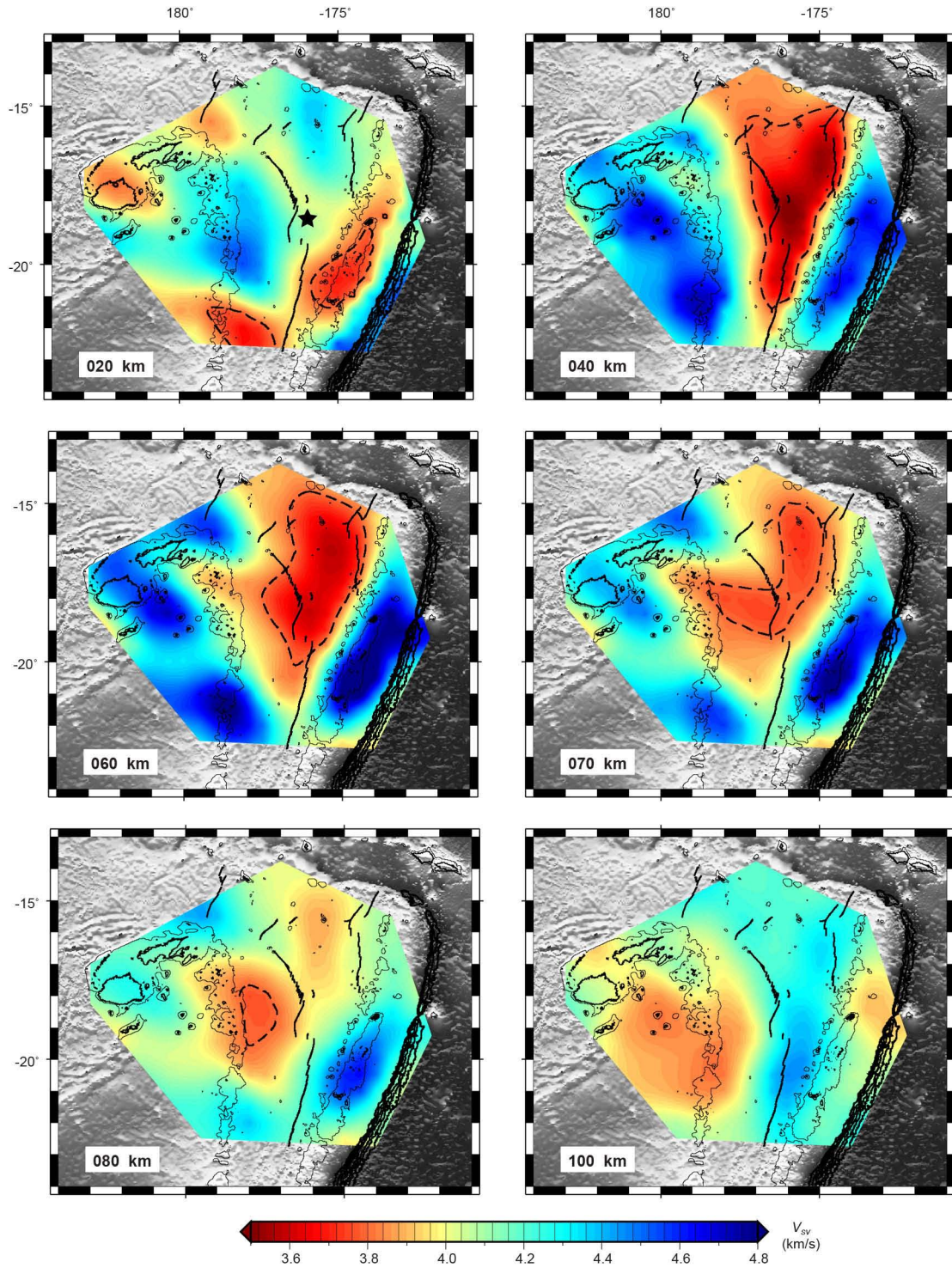
$$V(T) = V_0 \left[1 - \frac{1}{2Q(T)} \cot\left(\frac{\pi\alpha}{2}\right) \right]$$

Here V_0 is the velocity at infinite frequency and α is the frequency-dependent exponent. Assuming that attenuation doubles (Q^{-1} increases from 0.04 to 0.08) when temperature increases by 100 °C ($d(Q^{-1})/dT = 0.0004 \text{ K}^{-1}$; ref. 18), that reference $V_0 = 4.0 \text{ km s}^{-1}$ and that $dV_0/dT = 0.000378$ (ref. 56), we have

$$\frac{dV}{dT} = \left\{ \frac{dV_0}{dT} \left[1 - \frac{Q^{-1}}{2} \cot\left(\frac{\pi\alpha}{2}\right) \right] - \frac{d(Q^{-1})}{dT} \frac{V_0}{2} \cot\left(\frac{\pi\alpha}{2}\right) \right\} \\ \approx 0.002 \text{ km s}^{-1} \text{ K}^{-1}$$

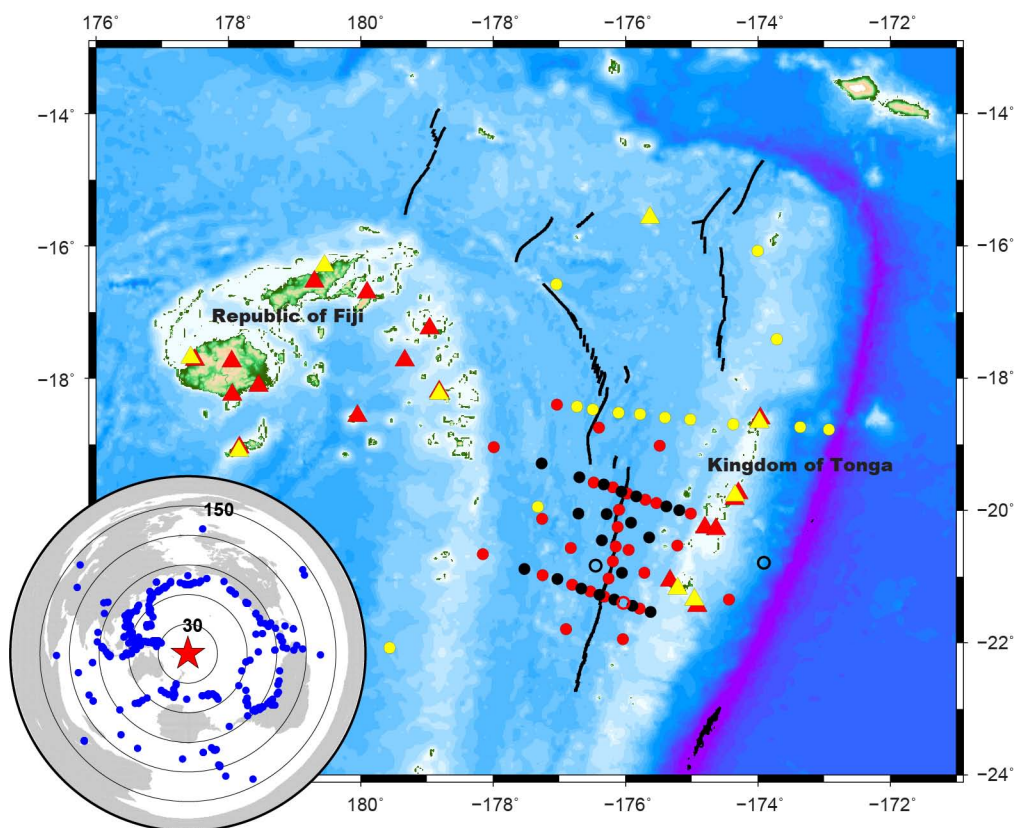
So the thermal difference of 50 °C between the CLSC and the VFR may result in a change of only 0.1 km s^{-1} in shear-wave velocity. Therefore, we suggest that the along-strike thermal variation is not sufficient to cause the observed change in seismic velocity, which requires a significant decrease of melt porosity towards the VFR.

30. Webb, S. C. & Crawford, W. C. Long-period seafloor seismology and deformation under ocean waves. *Bull. Seismol. Soc. Am.* **89**, 1535–1542 (1999).
31. Crawford, W. C. & Webb, S. C. Identifying and removing tilt noise from low-frequency (<0.1 Hz) seafloor vertical seismic data. *Bull. Seismol. Soc. Am.* **90**, 952–963 (2000).
32. Bell, S., Forsyth, D. W. & Ruan, Y. Removing noise from the vertical component records of ocean bottom seismometers: results from year one of the Cascadia Initiative. *Bull. Seismol. Soc. Am.* <http://dx.doi.org/10.1785/0120140054> (2015).
33. Smith, G. P. *et al.* A complex pattern of mantle flow in the Lau backarc. *Science* **292**, 713–716 (2001).
34. Herrmann, R. B. *Computer Programs in Seismology* v. 3.30 (Earthquake Center, St Louis Univ., 2004).
35. Wiens, D. A., Kelley, K. A. & Plank, T. Mantle temperature variations beneath back-arc spreading centers inferred from seismology, petrology, and bathymetry. *Earth Planet. Sci. Lett.* **248**, 30–42 (2006).
36. Crawford, W. C., Hildebrand, J. A., Dorman, L. M., Webb, S. C. & Wiens, D. A. Tonga Ridge and Lau Basin crustal structure from seismic refraction data. *J. Geophys. Res.* **108**, 2195 (2003).
37. Conder, J. A. & Wiens, D. A. Seismic structure beneath the Tonga arc and Lau back-arc basin determined from joint Vp, Vp/Vs tomography. *Geochem. Geophys. Geosyst.* **7**, Q03018 (2006).
38. Rau, C. J. & Forsyth, D. W. Melt in the mantle beneath the amagmatic zone, southern Nevada. *Geology* **39**, 975–978 (2011).
39. Pyle, M. L. *et al.* Shear velocity structure of the Mariana mantle wedge from Rayleigh wave phase velocities. *J. Geophys. Res.* **115**, B11304 (2010).
40. Shen, Y. & Forsyth, D. W. Geochemical constraints on initial and final depths of melting beneath mid-ocean ridges. *J. Geophys. Res.* **100**, 2211–2237 (1995).
41. Wasylenko, L. E., Baker, M. B., Kent, A. J. R. & Stolper, E. M. Near-solidus melting of the shallow upper mantle: partial melting experiments on depleted peridotite. *J. Petrol.* **44**, 1163–1191 (2003).
42. Lee, C.-T. A. Compositional variation of density and seismic velocities in natural peridotites at STP conditions: implications for seismic imaging of compositional heterogeneities in the upper mantle. *J. Geophys. Res.* **108**, 2441 (2003).
43. Lehnert, K., Su, Y., Langmuir, C. H., Sarbas, B. & Nohl, U. A global geochemical database structure for rocks. *Geochem. Geophys. Geosyst.* **1**, 1012 (2000).
44. Kamenetsky, V. S., Crawford, A. J., Eggins, S. & Mühe, R. Phenocryst and melt inclusion chemistry of near-axis seamounts, Valu Fa Ridge, Lau Basin: insight into mantle wedge melting and the addition of subduction components. *Earth Planet. Sci. Lett.* **151**, 205–223 (1997).
45. Kent, A. J. R., Peate, D. W., Newman, S., Stolper, E. M. & Pearce, J. A. Chlorine in submarine glasses from the Lau Basin: seawater contamination and constraints on the composition of slab-derived fluids. *Earth Planet. Sci. Lett.* **202**, 361–377 (2002).
46. Tian, L. *et al.* Major and trace element and Sr–Nd isotope signatures of lavas from the Central Lau Basin: implications for the nature and influence of subduction components in the back-arc mantle. *J. Volcanol. Geotherm. Res.* **178**, 657–670 (2008).
47. Hahn, D. *et al.* An overview of the volatile systematics of the Lau Basin: resolving the effects of source variation, magmatic degassing and crustal contamination. *Geochim. Cosmochim. Acta* **85**, 88–113 (2012).
48. Lytle, M. L. *et al.* Tracing mantle sources and Samoan influence in the northwestern Lau back-arc basin. *Geochem. Geophys. Geosyst.* **13**, Q10019 (2012).
49. Langmuir, C. H., Bézous, A., Escrig, S. & Parman, S. W. in *Back-Arc Spreading Systems: Geological, Biological, Chemical, and Physical Interactions* (eds Christie, D. M., Fisher, C. R., Lee, S.-M. & Givens, S.) 87–146 (American Geophysical Union, 2006).
50. Kelley, K. A. & Cottrell, E. Water and the oxidation state of subduction zone magmas. *Science* **325**, 605–607 (2009).
51. Hirschmann, M. M. Partial melt in the oceanic low velocity zone. *Phys. Earth Planet. Inter.* **179**, 60–71 (2010).
52. Tenner, T. J., Hirschmann, M. M. & Humayun, M. The effect of H₂O on partial melting of garnet peridotite at 3.5 GPa. *Geochem. Geophys. Geosyst.* **13**, Q03016 (2012).
53. Hirschmann, M. M., Tenner, T., Aubaud, C. & Withers, A. C. Dehydration melting of nominally anhydrous mantle: the primacy of partitioning. *Phys. Earth Planet. Inter.* **176**, 54–68 (2009).
54. Cooper, L. B. *et al.* High-Ca boninites from the active Tonga Arc. *J. Geophys. Res.* **115**, B10206 (2010).
55. Karato, S.-I. Importance of anelasticity in the interpretation of seismic tomography. *Geophys. Res. Lett.* **20**, 1623–1626 (1993).
56. Stixrude, L. & Lithgow-Bertelloni, C. Mineralogy and elasticity of the oceanic upper mantle: origin of the low-velocity zone. *J. Geophys. Res.* **110**, B03204 (2005).
57. Hirschmann, M. M. Mantle solidus: experimental constraints and the effects of peridotite composition. *Geochem. Geophys. Geosyst.* **1**, 1042 (2000).



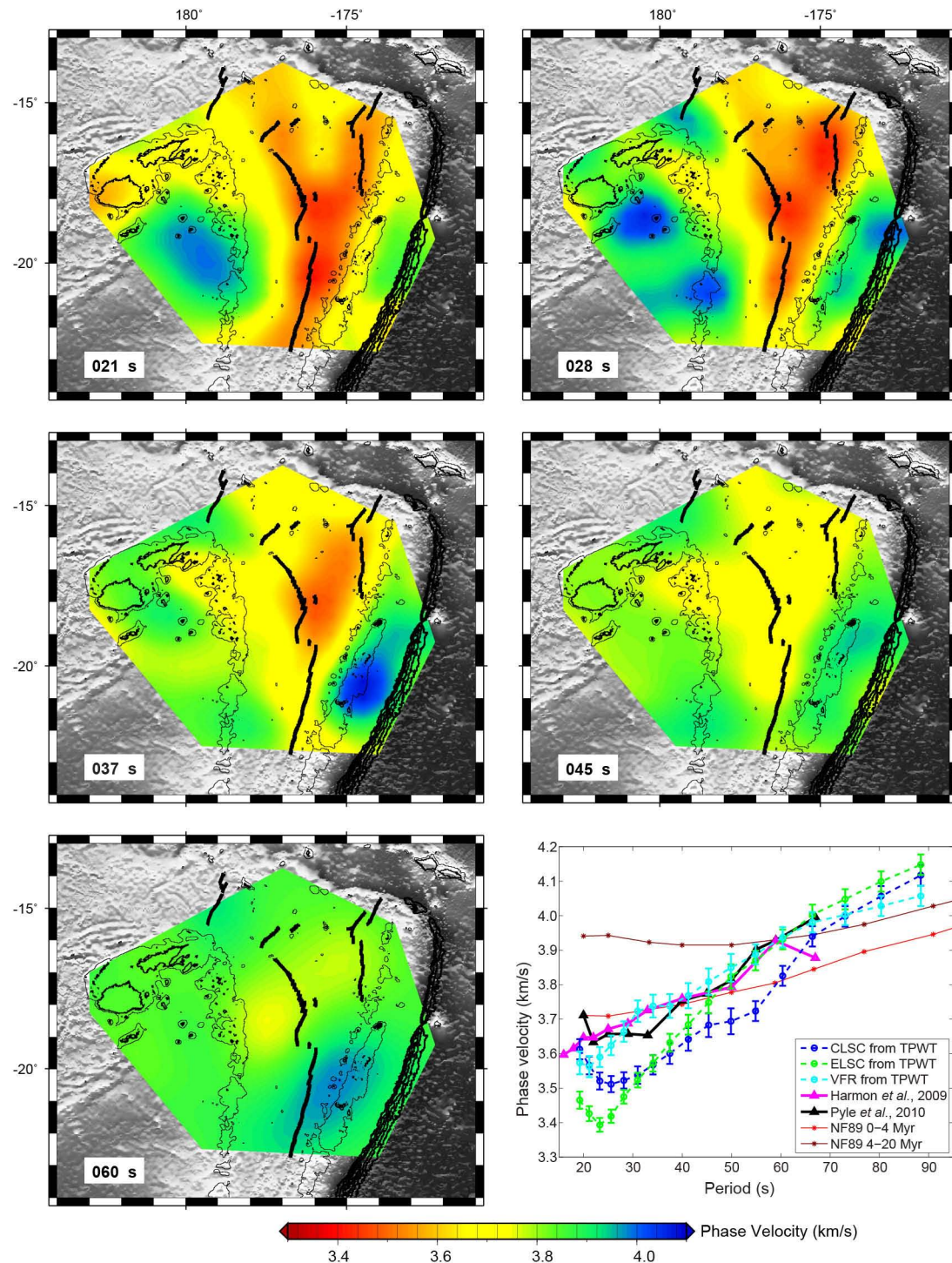
Extended Data Figure 1 | Maps of azimuthally averaged SV-wave velocity at depths of 20, 40, 60, 70, 80 and 100 km. S-wave velocity of 3.8 km s^{-1} is contoured. Star illustrates node 364, used in the Monte Carlo inversion

(Extended Data Fig. 7). Spreading centres and bathymetry contours are labelled as in Fig. 1c.



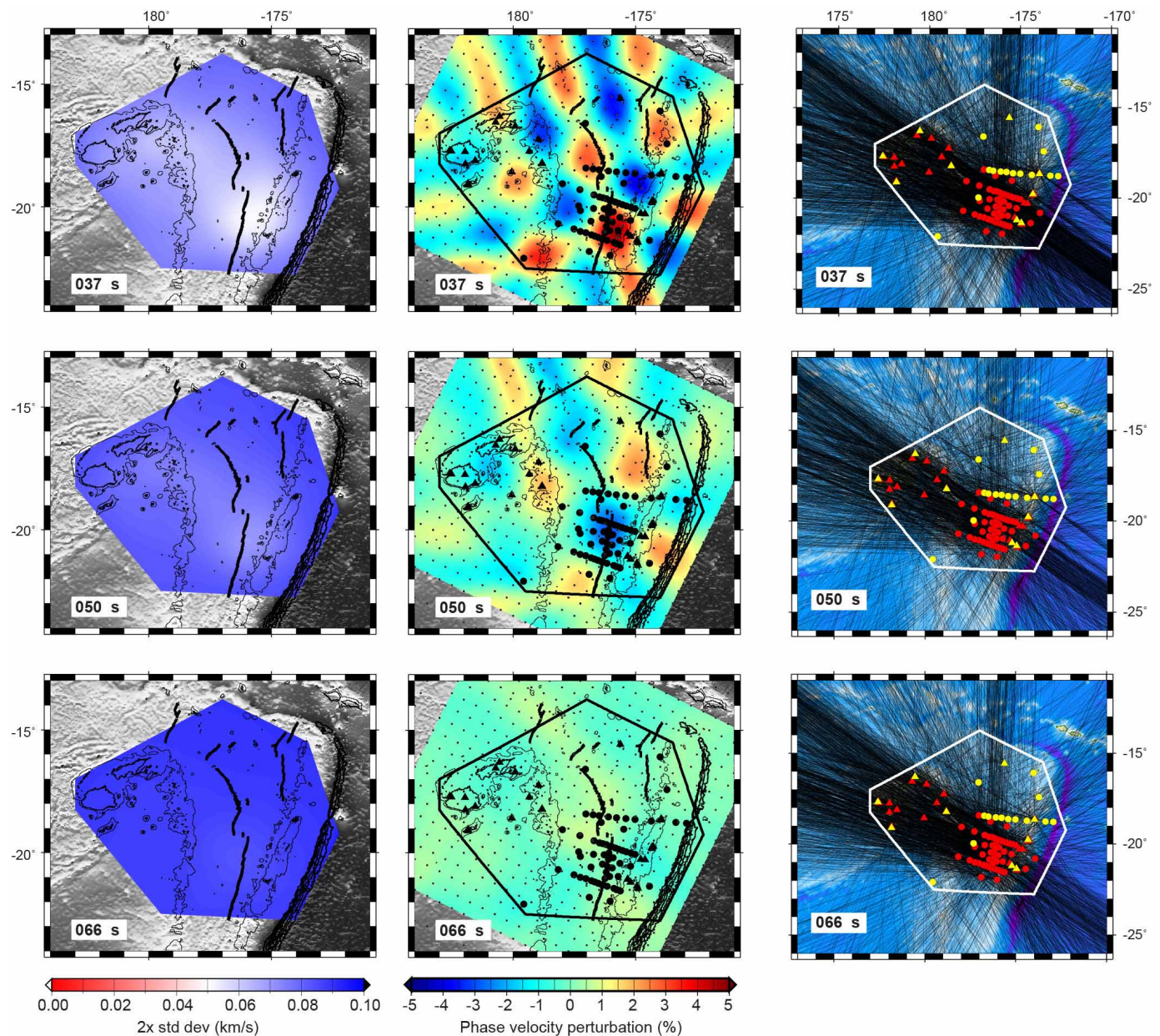
Extended Data Figure 2 | Seismic stations and earthquakes used in this study. Red triangles represent island-based stations operated from October 2009 to December 2010. Red and black dots are WHOI (Woods Hole Oceanographic Institution) and LDEO (Lamont-Doherty Earth Observatory) OBSs deployed from November 2009 to November 2010. Open circles mean

unrecovered OBSs. Yellow dots and triangles indicate OBSs and island-based stations deployed during September to December 1994, respectively. Spreading centres and bathymetry are labelled as in Fig. 1b. The inset shows the earthquakes (blue dots) used in this study centred at the Lau Basin (red star).



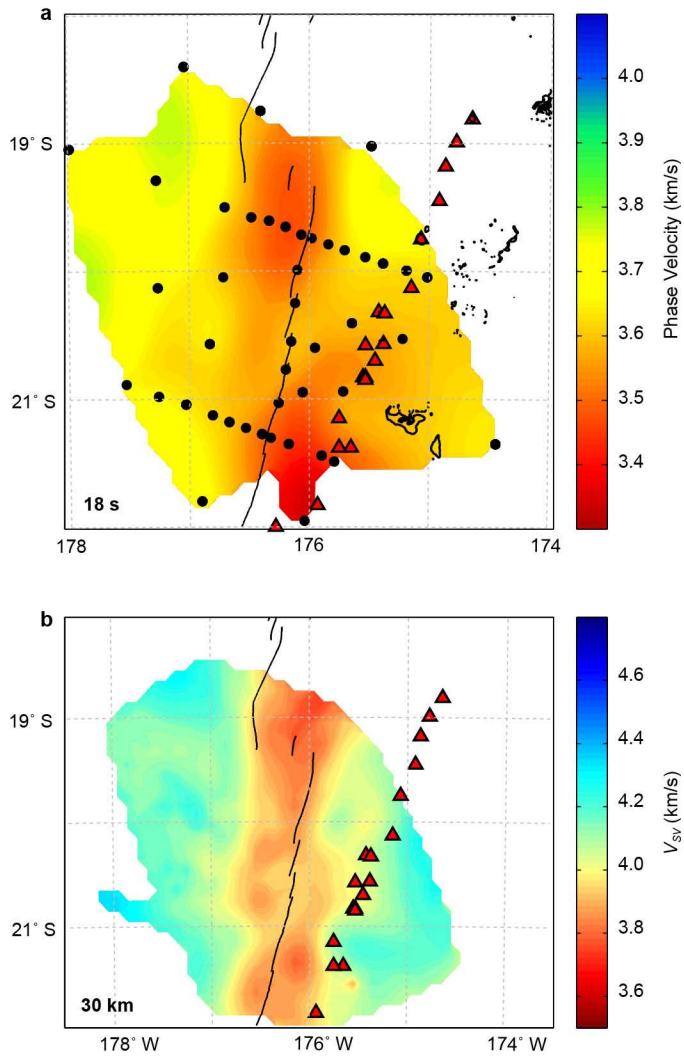
Extended Data Figure 3 | Maps of azimuthally isotropic phase velocity at periods of 21, 28, 37, 45 and 60 s inverted with the finest inverting grid. Spreading centres and bathymetry contours are labelled as in Fig. 1c. Dispersion curves are shown for the CLSC (blue), ELSC (green), VFR (cyan),

East Pacific Rise¹³ (magenta), Mariana back-arc³⁹ (black) and NF89 models¹⁴ (red and dark red). The CLSC, ELSC and VFR are represented by nodes shown in Fig. 1d. Error bars indicate the standard deviations of phase velocity.

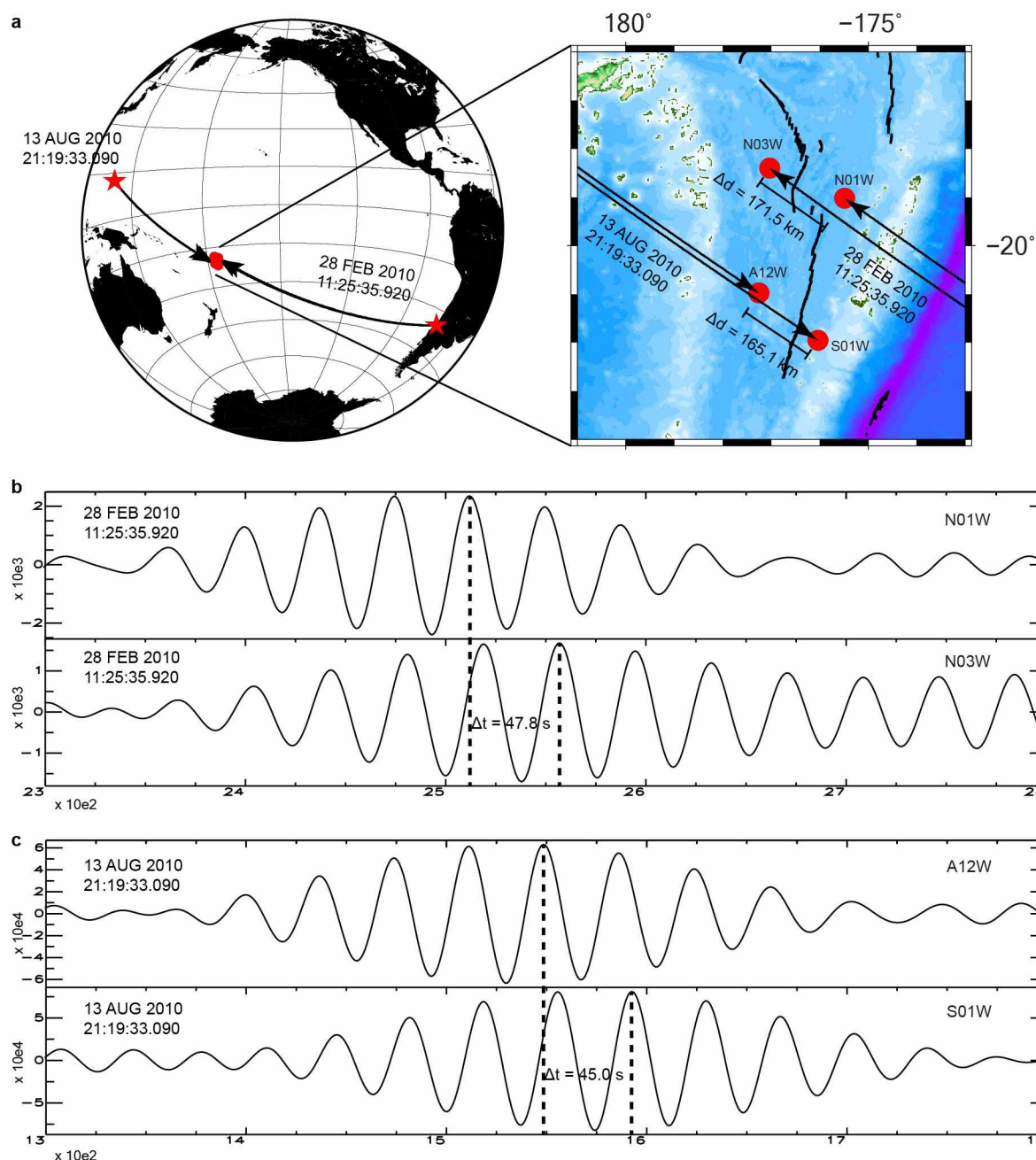


Extended Data Figure 4 | Robustness of the phase-velocity inversion at periods of 37, 50 and 66 s, which are most sensitive to depths of about 50 km (where the velocity is lowest), 70 km (where the inclined LVZ extends away from the trench) and 100 km (the maximum depth to be interpreted), respectively. Left panels: maps of double standard deviation inverted with the finest grid. Middle panels: resolution test of phase-velocity inversion with the finest inverting grid (regularly spacing black points). Black dots and

triangles represent 63 OBSs and 26 land-based seismic stations used in this study, respectively. The black polygon outlines the region in which we display results because within it we achieved a reasonable resolution of phase-velocity inversion at all periods. Spreading centres and bathymetry contours are labelled as in Fig. 1c. Right panels: Rayleigh wave ray-paths (black lines) used in phase-velocity inversion. Seismic stations are labelled as in Fig. 1b.

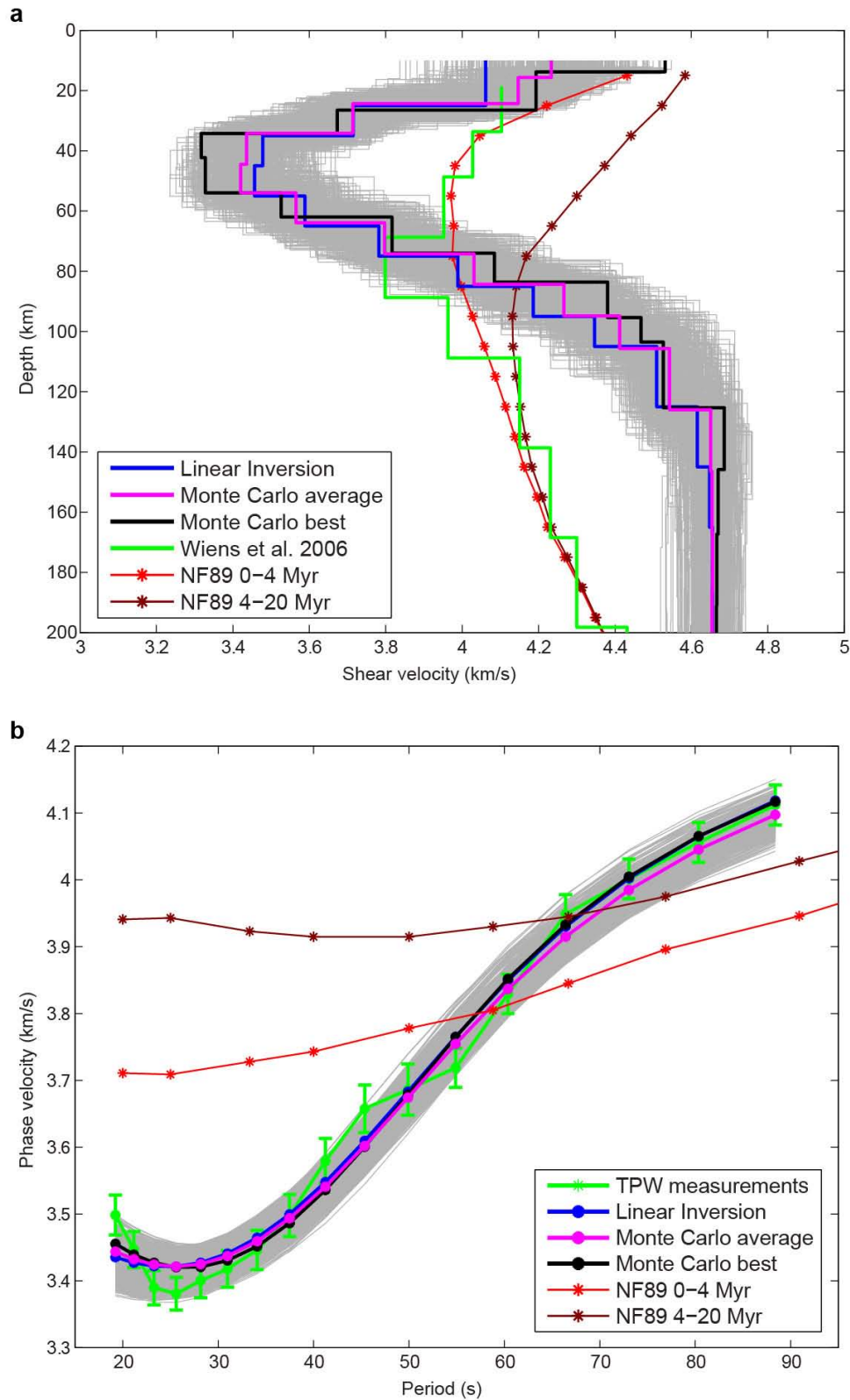


Extended Data Figure 5 | Previous results of ambient-noise tomography.
a, Isotropic phase velocities of Rayleigh waves at the period of 18 s. Black dots indicate the OBSs used in this study, red triangles represent active volcanoes, and black lines mean the spreading centres. **b**, Azimuthally averaged SV-wave velocity at a depth of 30 km. All ANT results from ref. 15.



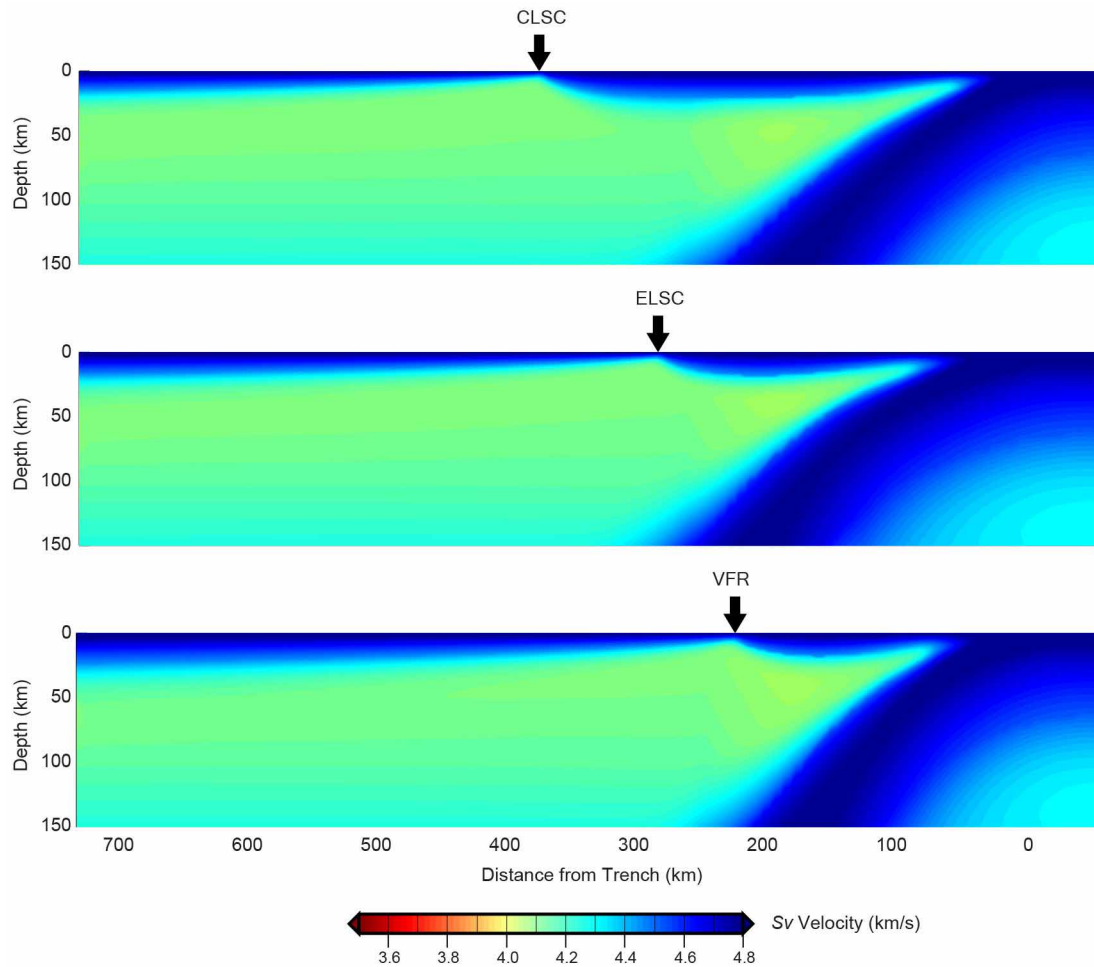
Extended Data Figure 6 | Two examples of phase velocity measured by the two-station method. **a**, Surface waves (black curves) of two earthquakes (red stars) propagated to four OBSs (red dots). **b**, The earthquake at the Chile trench was recorded by stations N01W and N03W. The difference in epicentral distances is about 171.5 km. The Rayleigh wave at a period of 37 s has

a delay time of 47.8 s, suggesting a phase velocity of 3.59 km s^{-1} . **c**, The earthquake at the Mariana trench was recorded by stations A12W and S01W. The difference in epicentral distances is about 165.1 km. The Rayleigh wave at a period of 37 s has a delay time of 45.0 s, suggesting a phase velocity of 3.67 km s^{-1} .



Extended Data Figure 7 | SV-wave-velocity inversion of Monte Carlo algorithm for node 364. **a**, Models of SV-wave velocity. **b**, Forward-calculated dispersion curves. Each grey curve indicates one 'good' model whose smoothness and mis-fit are smaller than the criteria. Blue, magenta, black, red and dark red curves represent the model from linearized inversion, the average

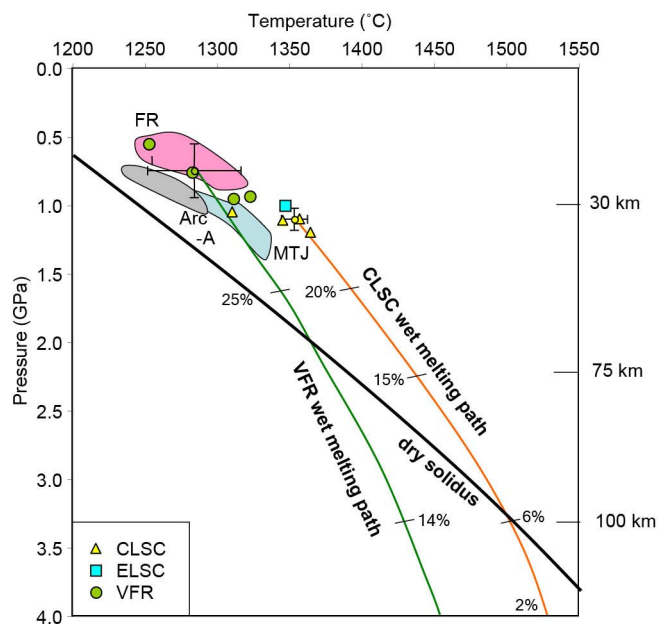
model, the best model from Monte Carlo inversion and NF89 models of two age categories¹⁴, respectively. In **a**, the model of SV-wave velocity from ref. 35 (green) is shown for reference. In **b**, the green curve indicates phase velocities inverted by TPWT, with error bars showing the standard deviations.



Extended Data Figure 8 | Cross-sections of predicted SV-wave velocity.

Calculations are based on numerical models of temperature and water content²³, the extended Burgers model²⁰, and corrections for radial anisotropy¹⁴ and effects of water²⁴. The colour scale is the same as in Fig. 2. Although,

compared with that beneath the CLSC, the temperature beneath the VFR is lower owing to slab cooling, which potentially increases the seismic velocity, the much higher water content reduces the velocity more significantly and leads to a stronger signal of low velocity in the prediction.



Extended Data Figure 9 | Pressures and temperatures of equilibration of Lau Basin glasses with Fo90 mantle. We used major elements, H_2O measurements, f_{O_2} constraints and the thermobarometer of ref. 27 to calculate P - T paths for most primitive melts (crystallizing olivine only). Back-arc averages for the VFR and the CLSC-ELSC high-temperature cluster are shown with smaller symbols and error bars of 1 s.d. Fields are given for the FRSC, the MTJ and Tonga Arc Volcano A for comparison. All data are from PetDB⁴³⁻⁴⁸ and ref. 54. The dry solidus is from ref. 57. Back-arc averages are traced back along wet decompression melting paths, as described in Methods.

Functional organization of excitatory synaptic strength in primary visual cortex

Lee Cossell^{1,2*}, Maria Florencia Iacaruso^{1,2*}, Dylan R. Muir², Rachael Houlton¹, Elie N. Sader¹, Ho Ko^{1,3}, Sonja B. Hofer^{1,2} & Thomas D. Mrsic-Flogel^{1,2}

The strength of synaptic connections fundamentally determines how neurons influence each other's firing. Excitatory connection amplitudes between pairs of cortical neurons vary over two orders of magnitude, comprising only very few strong connections among many weaker ones^{1–9}. Although this highly skewed distribution of connection strengths is observed in diverse cortical areas^{1–9}, its functional significance remains unknown: it is not clear how connection strength relates to neuronal response properties, nor how strong and weak inputs contribute to information processing in local microcircuits. Here we reveal that the strength of connections between layer 2/3 (L2/3) pyramidal neurons in mouse primary visual cortex (V1) obeys a simple rule—the few strong connections occur between neurons with most uncorrelated responses, while only weak connections link neurons with uncorrelated responses. Moreover, we show that strong and reciprocal connections occur between cells with similar spatial receptive field structure. Although weak connections far outnumber strong connections, each neuron receives the majority of its local excitation from a small number of strong inputs provided by the few neurons with similar responses to visual features. By dominating recurrent excitation, these infrequent yet powerful inputs disproportionately contribute to feature preference and selectivity. Therefore, our results show that the apparently complex organization of excitatory connection strength reflects the similarity of neuronal responses, and suggest that rare, strong connections mediate stimulus-specific response amplification in cortical microcircuits.

To determine the relationship between connection strength and neuronal responses, we used a combination of two-photon calcium imaging *in vivo* and whole-cell recordings *in vitro* in L2/3 of mouse V1¹⁰ (Fig. 1). We first examined how connection strength relates to the degree of correlated firing between pairs of neurons. We obtained pairwise correlation coefficients of responses to a sequence of static natural images (see Methods) from L2/3 neurons labelled with the calcium-sensitive indicator OGB-1 (ref. 11; imaged volumes $\sim 260 \times 260 \times 56 \mu\text{m}$; Fig. 1a, b). The distribution of pairwise response correlations was highly skewed: correlations were generally low, and only a small fraction of pairs were highly correlated during visual stimulation (median correlation coefficient: 0.012; mean correlation coefficient \pm s.d.: 0.021 ± 0.051 ; range: -0.12 to 0.67 ; Fig. 1c; Extended Data Fig. 1).

We next identified the same OGB-1-filled neurons in acute slices (Fig. 1e), and targeted up to six neurons for simultaneous whole-cell recording to assess their synaptic connectivity (Fig. 1e, f; see Methods). A total of 203 pyramidal cells (across 17 mice) recorded in the slice were identified in the *in vivo* image stacks, and the overall connection rate was 75/520 (0.14). Consistent with previous reports^{1–9}, the distribution of excitatory postsynaptic potential (EPSP) amplitudes was highly skewed (median EPSP amplitude: 0.19 mV; mean EPSP amplitude \pm s.d.: 0.45 ± 0.68 mV; Fig. 1d).

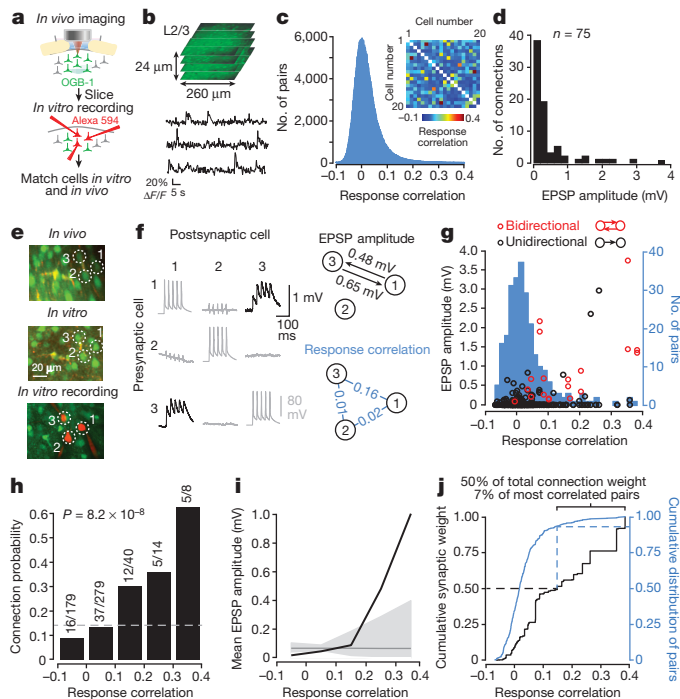
Neurons with more similar responses were much more likely to connect ($P = 8.2 \times 10^{-8}$, Cochran–Armitage test for trend; Fig. 1g, h), consistent with previous observations¹⁰. Importantly, response correlation was closely related to EPSP amplitude: the strongest connections were found between neuronal pairs with the highest response correlations (Fig. 1g, i; Extended Data Fig. 2a), while neurons with negatively correlated responses tended not to connect or formed only weak connections (response correlation > 0.2 , median amplitude of connected pairs: 1.38 mV, $n = 10$; response correlation < 0 , median amplitude: 0.12 mV, $n = 16$; $P = 6.8 \times 10^{-4}$, Wilcoxon rank-sum test; Fig. 1g). The close correspondence between response correlation and mean connection amplitude was apparent both when including (Fig. 1i) or excluding unconnected pairs (Extended Data Fig. 2a). Indeed, the majority of total synaptic weight was concentrated in the minority of connections between highly correlated pairs (7% most correlated pairs accounted for 50% of the total synaptic weight; Fig. 1j), further emphasizing the highly non-random arrangement of connection amplitudes. Together, these data suggest that the long-tailed distribution of cortical connection weights arises from a simple rule: neurons with highly correlated responses form strong connections, and neurons with uncorrelated responses connect rarely, and only weakly.

In visual cortex, strong response correlations may be explained by one of several shared visual response properties. To understand how connection strength relates to visual feature preference, we characterized the spatial linear receptive field (RF) structure for each neuron in the imaged populations (Fig. 2a; see Methods). The linear RF describes the relative position of ON (response to light increments) and OFF (response to light decrements) subfields in visual space, and thus provides information about visual features to which a neuron is most sensitive, including their orientation, phase, spatial frequency and size. RFs of nearby cortical neurons were highly diverse within each imaged population in L2/3^{12–15} (example region in Fig. 2b). We quantified RF similarity as the pixel-to-pixel correlation coefficient between pairs of RF maps. RF correlations were close to zero for the majority of pairs, and only a small fraction of neurons exhibited highly similar or highly dissimilar RFs (Fig. 2c).

By assessing connectivity between neurons with linear RFs (Fig. 2d–f), we found that connections between neuronal pairs with more similar RF structure were stronger (Fig. 2g, i; Extended Data Fig. 2b) and much more frequent (Fig. 2h) than connections between pairs with uncorrelated or negatively correlated RFs. The minority of connections observed between neurons with the most similar RFs accounted for a large fraction of the total synaptic weight (50% of total synaptic weight between 12% of pairs with the most correlated RFs; Fig. 2j). RF correlation was closely related to the degree of ON and OFF subfield overlap ($R = 0.79$, $P < 1 \times 10^{-10}$; Extended Data Fig. 3a), and both connection strength and connection probability increased with larger ON and OFF overlap (Extended Data Fig. 3b–d). Indeed, these measures of RF similarity predicted

¹Department of Neuroscience, Physiology and Pharmacology, University College London, 21 University Street, London WC1E 6DE, UK. ²Biozentrum, University of Basel, Klingelbergstrasse 50/70, CH-4056 Basel, Switzerland. ³Lui Che Woo Institute of Innovative Medicine and Chow Yuk Ho Technology Center for Innovative Medicine, Faculty of Medicine, the Chinese University of Hong Kong, Shatin, New Territories, Hong Kong.

*These authors contributed equally to this work.



connection strength much better than cortical distance and even the difference in orientation preference (Fig. 2k; Extended Data Fig. 4; see Methods). Nonetheless, pairwise response correlation was the best predictor of connection strength (Fig. 2k); this is not surprising as pairwise response correlation includes additional information about shared response properties not captured by the linear RF¹⁶.

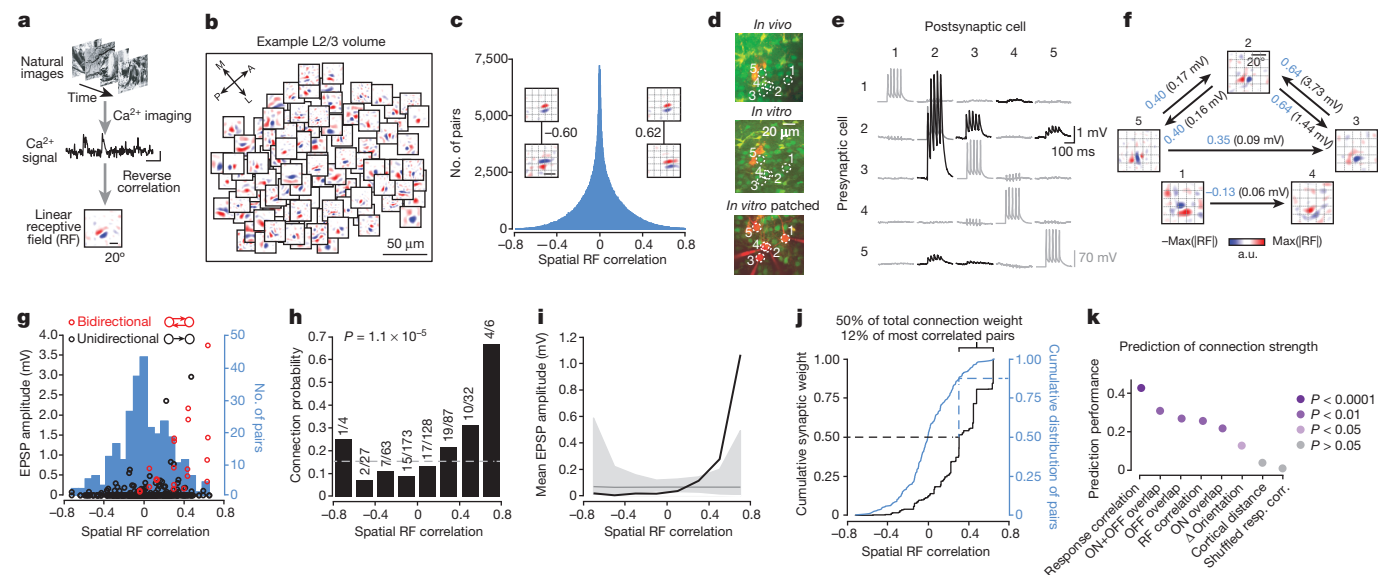


Figure 2 | Organization of excitatory connection strength with respect to linear RF properties. **a**, RFs obtained by regularized reverse correlation of responses to a sequence of static natural images (see Methods; scale bars: calcium trace, 20% $\Delta F/F$, 5 s, RF, 20°). **b**, RFs distributed across an example imaged region (collapsed over cortical depth, 150–174 μm below cortical surface), revealing a large diversity of RFs. **c**, Distribution of spatial RF correlation coefficients for all recorded cell pairs. Inset, example RFs for two pairs of neurons and their correlation coefficients. Typically, negatively correlated RFs had similar orientation but opposite phase preference. **d**, Example quintuplet of neurons shown in the transformed *in vivo* image (upper), in the brain slice (middle) and during whole-cell recordings (lower). **e**, Average postsynaptic potential traces of neurons in **d**. Black traces,

Figure 1 | Excitatory connection strength reflects the similarity of pyramidal cell firing *in vivo*. **a**, Schematic of experimental protocol. **b**, Somatic calcium signals were sampled simultaneously from all neurons within a small volume of cortex ($\sim 260 \times 260 \times 24 \mu\text{m}$). Two such volumes were recorded in each experiment. **c**, The distribution of pairwise response correlation coefficients for all imaged cell pairs. Inset, example matrix of correlation coefficients of pairwise responses from 20 neurons within a single imaged volume. **d**, Distribution of excitatory postsynaptic potential (EPSP) amplitudes ($n = 75$ connections). **e**, Example triplet of neurons shown in a transformed *in vivo* image (upper), in the brain slice (middle) and during whole-cell recordings (lower). **f**, Left, average postsynaptic potential traces of neurons in **e**. Black traces, connected; grey traces, unconnected; evoked presynaptic spikes are along the diagonal. In some traces, capacitive stimulation artefacts coincide with presynaptic spikes. Right, synaptic connectivity and response correlation coefficients of neurons in **e**. **g**, EPSP amplitude plotted against pairwise response correlation coefficient for bidirectionally (red) and unidirectionally (black) connected pairs. Underlying histogram shows the distribution of pairwise response correlation coefficients (blue, right y axis). **h**, Relationship between connection probability and pairwise response correlation. Grey dashed line, mean connection probability. Connection probability increased with response correlation ($P = 8.2 \times 10^{-8}$; Cochran–Armitage test). **i**, Mean connection amplitude (including unconnected pairs) plotted against response correlation (bin size = 0.1). Grey line, mean EPSP amplitude of all pairs. Grey shaded region represents the 95% confidence interval of the expected mean, estimated by repeated random reshuffling of the EPSP amplitudes among all cell pairs in the data set. **j**, Black trace, cumulative distribution of synaptic weight with respect to response correlation. A value of 1 corresponds to the linear sum of all EPSPs (33.34 mV). Blue trace, cumulative distribution of pairwise response correlation coefficients (right y axis).

The preferential, strong connectivity between correlated neurons was further emphasized when considering the reciprocity of connections (Figs 1g and 2g, Extended Data Fig. 5). Connections between bidirectionally connected pairs generated larger EPSPs than unidirectionally connected pairs, consistent with previous reports^{3,4,17}, and the RF maps of bidirectionally connected neurons were significantly more correlated

connected; grey traces, unconnected; evoked presynaptic spikes are along the diagonal. **f**, Synaptic connectivity and RFs of neurons in **d**. Arrows indicate a synaptic connection. Values indicate the correlation coefficient of RF maps (blue) and the amplitude of the connection (EPSP, black). a.u., arbitrary units. **g**, EPSP amplitude plotted against RF correlation for bidirectionally (red) and unidirectionally (black) connected pairs. Underlying histogram shows the distribution of pairwise RF correlations (blue, right y axis). **h–j**, Same as Fig. 1h–j for the RF correlation coefficient. **k**, Similarity of shared neuronal properties ranked according to how well they predict connection amplitude (including unconnected pairs). Prediction performance and P values were calculated using a Monte-Carlo analysis (see Methods). Disc colour indicates P value.

than RFs of unidirectionally connected or unconnected pairs (Extended Data Fig. 5).

Only a small fraction of neuronal pairs in the local V1 network shared a similar RF structure (7.5% of pairs with RF correlation > 0.4 ; Fig. 2j). Thus, connections between neurons with non-matching RFs (RF correlation < 0.4 , 61/75 or 81% of all measured connections) greatly outnumbered the connections between neurons with similar RFs (RF correlation > 0.4 , 14/75 or 19% of all measured connections). Given this large RF diversity of local inputs, we next sought to estimate the combined visual feature preference of the net synaptic excitation an individual neuron receives from the L2/3 network. We combined data from all pairs of connected neurons after rotating, translating and scaling each postsynaptic RF to match a normalized RF structure¹⁸ (Extended Data Fig. 6a; see Methods). The same transformation was applied to the RFs of the presynaptic neurons (Extended Data Fig. 6b), and we considered the sum of these transformed presynaptic RFs, weighted by the amplitude of the connections, to indicate net synaptic input. The structure of this weighted presynaptic RF sum closely resembled the structure of the normalized postsynaptic RF (correlation between presynaptic RF sum and postsynaptic RF sum: $R = 0.73$, $n = 45$ pairs; Fig. 3a, top row; Extended Data Fig. 7). This input specificity did not result from a bias in the structure of RFs in the local population because the RF sum of the unconnected neurons was very different ($R = -0.38$, $n = 227$; Fig. 3a, bottom row; Extended Data Fig. 8). Therefore, despite the majority of inputs arising from neurons with mismatched RFs, the combined local excitatory drive onto pyramidal cells in L2/3 has a RF structure that closely matches that of the receiving neuron.

How do connections of different strengths contribute to this feature-specific excitation from the local network? The RFs of the strongest 25% of inputs ($n = 11$), which accounted for 78% of the overall synaptic weight, were highly similar to the postsynaptic RF (correlation of weighted RF sums, pre- versus postsynaptic, $R = 0.67$; Fig. 3b). Further, the ON and OFF subfields of the strongest inputs closely matched the ON and OFF subfields of the postsynaptic RF, with very few mismatches (Fig. 3c, top row). Both the similarity between summed pre- and postsynaptic RFs (Fig. 3b), as well as the ON and OFF subfield overlap (Fig. 3c, middle rows), progressively decreased for weaker connections, and there was little common structure to the input from the weakest 25% of connections (Fig. 3c, bottom row). This indicates that feature-matched excitation from the local L2/3 network is dominated by only a small fraction of strong inputs that a pyramidal neuron receives.

We next sought to understand how local excitatory inputs contribute to the response and stimulus selectivity of a neuron's membrane potential in L2/3. In mouse V1, a simple cell's subthreshold response to drifting grating stimuli^{19–23} is characterized by two components that are determined by the angle and phase of the grating in relation to its RF (Fig. 4a). Namely, a large amplitude depolarization evoked at all orientations (F0 component, Fig. 4a–d), and an orientation-tuned membrane potential modulation locked to the grating phase (F1 component). The F1 component contributes directly to the firing response of a neuron, since spikes occur at the peaks of the large modulation (Fig. 4b). To estimate the contribution of local inputs to the F0 and F1 components, we generated a model informed by the experimentally measured RF properties and connectivity within L2/3 (see Methods). In the model network, a single neuron received input from all others in the population, and the input connection strength was drawn from the experimentally determined distribution relating RF correlation to connection strength (Fig. 4e, f). Despite receiving no direct feedforward input, simulated neurons displayed qualitatively similar membrane potential responses to L2/3 neurons recorded *in vivo* (Fig. 4g, compare to Fig. 4c), including a large depolarization (F0) at all stimulus orientations (Fig. 4g, h), and a highly modulated membrane potential response (F1) to the preferred but not the non-preferred orientation (Fig. 4g, h). These results suggest that local connections contribute to the stimulus selectivity of L2/3 neurons by providing tuned excitation that modulates the membrane potential in a manner qualitatively similar to that observed *in vivo*^{19–23}.

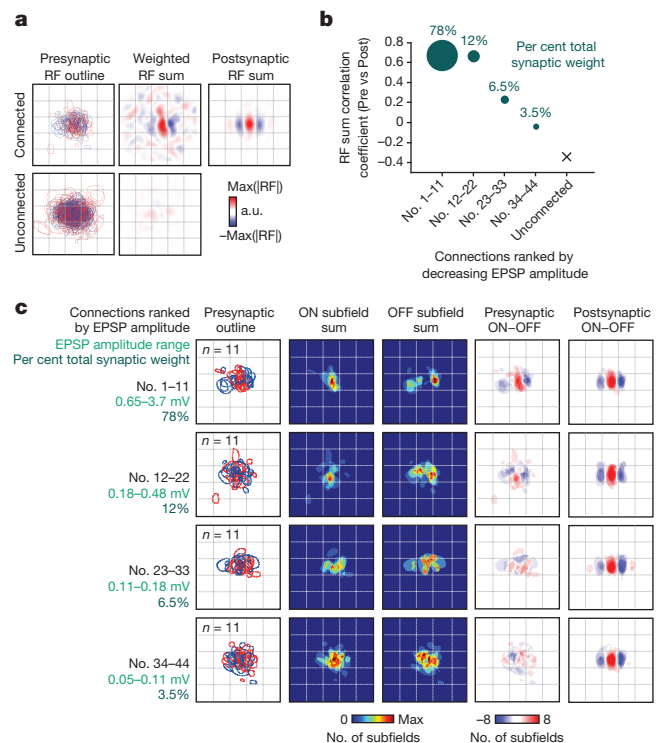


Figure 3 | Combined synaptic input from the local L2/3 cortical network matches the RF structure of the receiving neuron. **a**, Top left, presynaptic RF outlines overlaid in normalized visual space (after rotation, translation and scaling of the postsynaptic RF; see text, Methods and Extended Data Fig. 7). Red outline indicates an ON subfield, blue outline indicates an OFF subfield. Bottom left, superimposed RF outlines for neurons assessed presynaptically, but which did not connect. Top middle, sum of presynaptic RFs. Each presynaptic RF was weighted by the EPSP amplitude from the pre- to the postsynaptic neuron. Bottom middle, RF sum for unconnected neurons assessed presynaptically. Top right, RF sum of the postsynaptic neurons. Before summing, each postsynaptic RF was weighted by the EPSP amplitude from the presynaptic to the postsynaptic neuron. **b**, Each point indicates the correlation between the presynaptic RF sum (weighted by the EPSP amplitude) and the corresponding postsynaptic RF sum, when including only connections in quarters of the connection amplitude distribution. The RF sum of the strongest 25% of inputs has the highest correlation with the postsynaptic RF ($R = 0.67$). This correlation value falls with decreasing connection strength. Disc area and values above represent the total synaptic weight accounted for by each quarter of the connection amplitude distribution. **c**, Relationship between ON and OFF subfields of connected pre- and postsynaptic neurons, ranked and displayed according to EPSP amplitude. Left column, presynaptic RF outlines of neurons grouped in quarters ranked by decreasing EPSP amplitude. Middle-left and middle columns, sum of binarized presynaptic ON and OFF subfields, respectively. Middle-right column, subtraction of summed OFF from summed ON subfields for presynaptic neurons. Right column, subtraction of summed OFF from summed ON subfields for postsynaptic neurons.

We then systematically varied the relative fractions of strong and weak connections in the simulated network to estimate how connections of different strengths contribute to the F0 and F1 response components (Extended Data Fig. 9). For instance, removing the strongest 25% of connections from the model (equivalent to including the weakest 75% of connections; Fig. 4i, j, purple traces) essentially eliminated the modulated F1 component of the response (95% reduction in the modulation amplitude at the preferred orientation). In contrast, removing the weakest 75% of connections (leaving the strongest 25% connections; Fig. 4i, j, blue traces) decreased the mean depolarization (F0 component) by 23%, but only slightly affected the F1 component (5% decrease in modulation amplitude at the preferred orientation). This means that the membrane potential modulation of a L2/3 neuron in response to drifting gratings is weakly influenced by the majority of local inputs it receives,

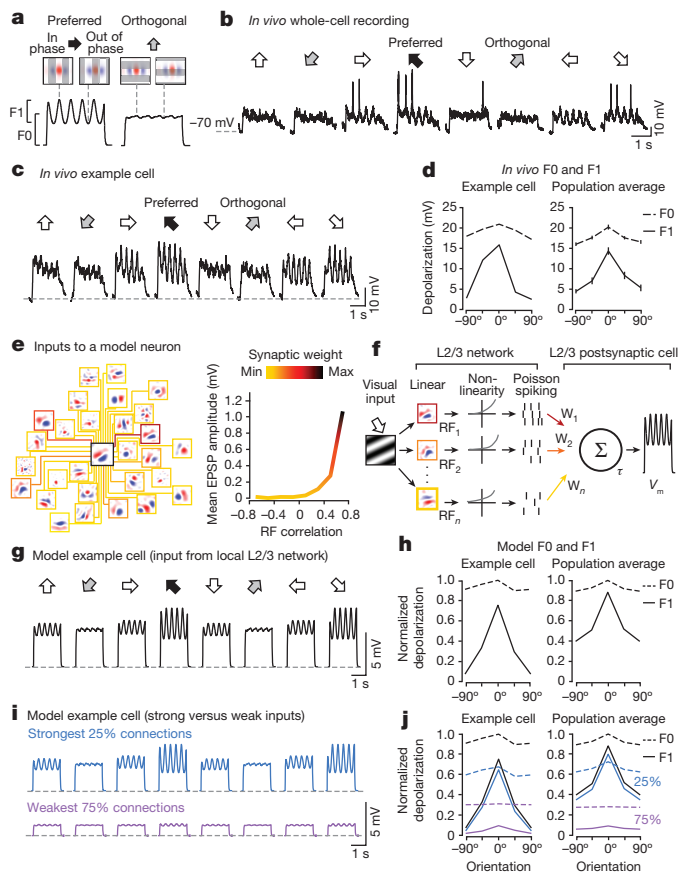


Figure 4 | Simulation of local L2/3 excitatory input to single neurons qualitatively predicts the dynamics of membrane depolarization to drifting grating stimuli. **a**, Schematic of characteristic membrane potential (V_m) response to gratings drifting across a neuron's RF in mouse V1. Both the preferred and the orthogonal stimuli evoke large membrane depolarizations (F0). V_m modulation (F1) is strongest when the grating and the RF are matched in orientation, and the grating cycles in and out of phase with the RF.

b, Example *in vivo* whole-cell V_m recording from a L2/3 pyramidal cell during presentation of oriented gratings drifting in eight different directions. Black and grey arrows indicate preferred and orthogonal orientations, respectively. **c**, Average V_m response of the same neuron in **b** after spike removal. **d**, F0 and F1 components of V_m response to drifting gratings normalized to the preferred stimulus for neuron in **b** and **c** (left), and averaged across the population of recorded neurons ($n = 24$, right). **e**, Left, schematic of network model, showing input to a single L2/3 neuron from an example L2/3 neuronal population. Right, connection strengths were sampled from the experimentally measured relationship between EPSP amplitude and RF correlation (Fig. 2g). **f**, Responses for each presynaptic neuron were generated using a linear/nonlinear/Poisson model by correlating the visual input (drifting grating stimuli) with its experimentally measured RF (see Methods). Firing rates were weighted by their connection strengths (**e**) and summed to generate a time-varying V_m for each postsynaptic neuron. **g**, Example V_m response of a simulated neuron receiving input from the model L2/3 network. **h**, F0 and F1 components of the V_m response from the example simulated neuron in **g** (left) or from the population of simulated neurons (right, $n = 4,633$). **i**, V_m response of example simulated neuron in **g** when including only the strongest 25% of connections (top, blue trace) or weakest 75% of connections from the model network (bottom, purple trace). **j**, Same as **d** and **h** but for V_m responses driven by the strongest 25% of connections (blue) or weakest 75% of connections (purple).

which contribute only partly to the broadly-tuned depolarization. In contrast, the orientation- and phase-selective response is predominantly influenced by a small subset of other L2/3 neurons with similar RF structure, which provide strong, feature-matched excitation at the preferred orientation.

We describe a simple rule governing how the long-tailed distribution of excitatory connection strength — observed in diverse cortical areas^{1–9}

— is organized with respect to the functional properties of cortical neurons. In L2/3 of mouse V1, pyramidal neurons with correlated responses to visual stimuli connect preferentially with strong and often reciprocal connections, whereas neurons with uncorrelated or anti-correlated responses connect infrequently with weak connections. The fact that there are only very few strong connections in neocortical circuits reflects the low number of cell pairs with highly correlated responses. Therefore, the strength of synaptic coupling mirrors the strength of functional coupling, a relationship which may arise from correlation-based learning rules^{15,24}.

Our results suggest that infrequent, strong connections play a prominent role in cortical computation. While a L2/3 pyramidal neuron receives inputs from many neurons with diverse response properties (for example, receptive fields) in the local V1 network, the majority of the synaptic drive is provided by only a small fraction of strong inputs from cells with the most similar responses to visual stimuli. These rare but powerful inputs provide strongly tuned excitation, and therefore directly contribute to a neuron's selectivity by amplifying responses to specific visual features.

This circuit architecture—comprising strong recurrent excitation within ensembles of neurons with similar RFs—may additionally amplify (and perhaps prolong²⁵) population-level responses to particular sensory stimuli^{19,20,26–29}, and thus promote effective information transmission to multiple postsynaptic targets. In contrast, the matrix of more numerous, weaker connections, which only generate a small fraction of total excitation in the L2/3 network, may facilitate local contextual interactions and serve as a substrate for plasticity—for example, when particular visual feature combinations become behaviourally relevant.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 10 October 2014; accepted 2 January 2015.

Published online 4 February 2015.

- Markram, H., Lübke, J., Frotscher, M., Roth, A. & Sakmann, B. Physiology and anatomy of synaptic connections between thick tufted pyramidal neurones in the developing rat neocortex. *J. Physiol. (Lond.)* **500**, 409–440 (1997).
- Feldmeyer, D., Lübke, J. & Sakmann, B. Efficacy and connectivity of intracolumnar pairs of layer 2/3 pyramidal cells in the barrel cortex of juvenile rats. *J. Physiol. (Lond.)* **575**, 583–602 (2006).
- Holmgren, C., Harkany, T., Svennerfors, B. & Zilberter, Y. Pyramidal cell communication within local networks in layer 2/3 of rat neocortex. *J. Physiol. (Lond.)* **551**, 139–153 (2003).
- Song, S., Sjöström, P. J., Reigl, M., Nelson, S. & Chklovskii, D. B. Highly nonrandom features of synaptic connectivity in local cortical circuits. *PLoS Biol.* **3**, e68 (2005).
- Lefort, S., Tomm, C., Floyd Sarria, J. & Petersen, C. H. The excitatory neuronal network of the C2 barrel column in mouse primary somatosensory cortex. *Neuron* **61**, 301–316 (2009).
- Buzsáki, G. & Mizuseki, K. The log-dynamic brain: how skewed distributions affect network operations. *Nature Rev. Neurosci.* **15**, 264–278 (2014).
- Morishima, M., Morita, K., Kubota, Y. & Kawaguchi, Y. Highly differentiated projection-specific cortical subnetworks. *J. Neurosci.* **31**, 10380–10391 (2011).
- Levy, R. B. & Reyes, A. D. Spatial profile of excitatory and inhibitory synaptic connectivity in mouse primary auditory cortex. *J. Neurosci.* **32**, 5609–5619 (2012).
- Tarczy-Hornoch, K., Martin, K. A. C., Stratford, K. J. & Jack, J. J. Intracortical excitation of spiny neurons in layer 4 of cat striate cortex *in vitro*. *Cereb. Cortex* **9**, 833–843 (1999).
- Ko, H. *et al.* Functional specificity of local synaptic connections in neocortical networks. *Nature* **473**, 87–91 (2011).
- Stosiek, C., Garaschuk, O., Holthoff, K. & Konnerth, A. *In vivo* two-photon calcium imaging of neuronal networks. *Proc. Natl Acad. Sci. USA* **100**, 7319–7324 (2003).
- Smith, S. L. & Häusser, M. Parallel processing of visual space by neighboring neurons in mouse visual cortex. *Nature Neurosci.* **13**, 1144–1149 (2010).
- Bonin, V., Histed, M. H., Yurgenson, S. & Reid, R. C. Local diversity and fine-scale organization of receptive fields in mouse visual cortex. *J. Neurosci.* **31**, 18506–18521 (2011).
- Niell, C. M. & Stryker, M. P. Highly selective receptive fields in mouse visual cortex. *J. Neurosci.* **28**, 7520–7536 (2008).
- Ko, H. *et al.* The emergence of functional microcircuits in visual cortex. *Nature* **496**, 96–100 (2013).
- Carandini, M. *et al.* Do we know what the early visual system does? *J. Neurosci.* **25**, 10577–10597 (2005).
- Perin, R., Berger, T. K. & Markram, H. A synaptic organizing principle for cortical neuronal groups. *Proc. Natl Acad. Sci. USA* **108**, 5419–5424 (2011).
- Reid, R. C. & Alonso, J. M. Specificity of monosynaptic connections from thalamus to visual cortex. *Nature* **378**, 281–284 (1995).

19. Lien, A. D. & Scanziani, M. Tuned thalamic excitation is amplified by visual cortical circuits. *Nature Neurosci.* **16**, 1315–1323 (2013).
20. Li, Y.-t., Ibrahim, L. A., Liu, B.-h., Zhang, L. I. & Tao, H. W. Linear transformation of thalamocortical input by intracortical excitation. *Nature Neurosci.* **16**, 1324–1330 (2013).
21. Tan, A. Y. Y., Brown, B. D., Scholl, B., Mohanty, D. & Priebe, N. J. Orientation selectivity of synaptic input to neurons in mouse and cat primary visual cortex. *J. Neurosci.* **31**, 12339–12350 (2011).
22. Jia, H., Rochefort, N. L., Chen, X. & Konnerth, A. Dendritic organization of sensory input to cortical neurons *in vivo*. *Nature* **464**, 1307–1312 (2010).
23. Smith, S. L., Smith, I., Branco, T. & Häusser, M. Dendritic spikes enhance stimulus selectivity in cortical neurons *in vivo*. *Nature* **503**, 115–120 (2013).
24. Clopath, C., Büsing, L., Vasilaki, E. & Gerstner, W. Connectivity reflects coding: a model of voltage-based STDP with homeostasis. *Nature Neurosci.* **13**, 344–352 (2010).
25. Druckmann, S. & Chklovskii, D. B. Neuronal circuits underlying persistent representations despite time varying activity. *Curr. Biol.* **22**, 2095–2103 (2012).
26. Li, L.-y., Li, Y.-t., Zhou, M., Tao, H. W. & Zhang, L. I. Intracortical multiplication of thalamocortical signals in mouse auditory cortex. *Nature Neurosci.* **16**, 1179–1181 (2013).
27. Douglas, R. J., Koch, C., Mahowald, M., Martin, K. A. & Suarez, H. H. Recurrent excitation in neocortical circuits. *Science* **269**, 981–985 (1995).
28. Ben-Yishai, R., Bar-Or, R. L. & Sompolinsky, H. Theory of orientation tuning in visual cortex. *Proc. Natl Acad. Sci. USA* **92**, 3844–3848 (1995).
29. Somers, D. C., Nelson, S. B. & Sur, M. An emergent model of orientation selectivity in cat visual cortical simple cells. *J. Neurosci.* **15**, 5448–5465 (1995).

Acknowledgements We thank K. Harris, G. Keller, T. Margrie, J. Sjöström, P. Znamenskiy and our laboratory members for discussions and comments, and T. Margrie, E. Rancz and J. Poulet for advice on *in vivo* whole-cell recordings. This work was supported by the Wellcome Trust (grant no. 095074) and the European Research council. L.C. was funded by the 4-year PhD programme in Neuroscience at UCL. M.F.I. was funded by a UCL International PhD fellowship. D.R.M. was supported by the University of Basel Young Researchers fund.

Author Contributions L.C., M.F.I., S.B.H. and T.D.M.-F. designed the experiments. L.C. and M.F.I. performed the *in vivo* and *in vitro* experiments. R.H. and E.N.S. performed the *in vivo* whole-cell recordings. H.K. and S.B.H. contributed to the transition during *in vivo* and *in vitro* experiments. H. K. and L.C. wrote the software for whole-cell recordings *in vivo* and *in vitro*. L.C., M.F.I. and D.R.M. analysed the data. L.C., M.F.I., D.R.M. and T.D.M.-F. wrote the manuscript. All authors discussed the data and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to T.D.M.-F. (thomas.mrsic-flogel@unibas.ch).

METHODS

Animals and surgical procedures. All experimental procedures were carried out in accordance with institutional animal welfare guidelines, and licensed by the UK Home Office.

***In vivo* two-photon calcium imaging.** Experiments were performed in 17 C57BL/6 mice of both sexes, aged P22–26. Please note that we applied no randomization or blinded assignment of animals to groups and no animals were excluded from any analyses. The number of experimental preparations used in the analysis was chosen according to previous studies^{10,13}. Mice were initially anaesthetized with a mixture of fentanyl (0.05 mg per kg), midazolam (5.0 mg per kg), and medetomidine (0.5 mg per kg). Surgery was performed as described previously³⁰. Briefly, we made a small craniotomy (1–2 mm) over primary visual cortex and, after dye injection, sealed it with 1.6% agarose in HEPES-buffered artificial cerebrospinal fluid (ACSF) as well as a cover slip. At the time of imaging, the injectable anaesthetic had mostly worn off and light anaesthesia was maintained by isoflurane (0.3–0.5%) in a 60:40% mixture of O₂:N₂O delivered via a small nose cone.

For bulk loading of cortical neurons, we first dissolved the calcium-sensitive dye Oregon Green BAPTA-1 acetyloxymethyl ester (OGB-1 AM, Molecular Probes) in 4 μ l DMSO containing 20% Pluronic F-127 (Molecular Probes), and then diluted (1/11) in dye buffer (150 mM NaCl, 2.5 mM KCl, and 10 mM HEPES, pH 7.4), resulting in a final concentration of 0.9 mM. In order to distinguish neurons and astrocytes, we added sulforhodamine 101 (SR 101, 50 μ M, Molecular Probes) to the solution³¹. With a micropipette (3–5 M Ω) the dye was slowly pressure-injected (3–10 psi, 2–4 min) into the monocular region of right visual cortex at a depth of 170–200 μ m under visual control by two-photon imaging ($\times 16$, 0.8 numerical aperture water immersion objective, Nikon). Activity of cortical neurons was monitored by two-photon imaging of OGB-1 fluorescence changes with a B-Scope microscope (Thorlabs) and a mode-locked Ti:sapphire laser (Mai Tai, Spectra-Physics) at 830 nm through a $\times 40$, 0.8 numerical aperture water immersion objective (Olympus).

Visual stimuli were generated using MATLAB (Mathworks) Psychophysics Toolbox^{32,33}, and displayed on an LCD monitor (60 Hz refresh rate) positioned 20 cm from the left eye, roughly at 45 degrees to the long axis of the animal, covering $\sim 110 \times 84$ degrees of visual space. At the beginning of each experiment, the appropriate retinotopic position in visual cortex was determined using small grating stimuli at 12 positions arranged in a 4×3 grid. The monitor was repositioned such that the preferred retinotopic position of most imaged neurons was roughly in the middle of the monitor.

Imaging frames of 512×512 pixels were acquired at 27.9 Hz in Scanimage 4.0 while presenting different visual stimuli, including movies and static naturalistic images (see sections below for details). A piezo *z*-scanner (PI P-726 PIFOC) was used to rapidly move the objective in the *z* axis and acquire 4 image planes, each separated by 8 μ m in depth. Thus a single imaged plane (acquired in 35.8 ms) was part of an imaged volume acquired in 143 ms (6.98 Hz). Before recording the first volume, a reference image was acquired. After each recording, the imaging position in the *x* and *y* axis was checked and realigned with the initial image if necessary. To obtain visually evoked responses from all neurons in a cortical volume of $263 \times 255 \times 56 \mu$ m, two volumes were recorded, starting at $\sim 135 \mu$ m below the cortical surface, corresponding to superficial layer 2 in mouse V1.

After aligning image sequences to correct for tangential drift we analysed them with customized programs written in MATLAB. A semi-automated algorithm was used to detect cell outlines, which were subsequently confirmed by visual inspection. This algorithm was based on morphological measurements of cell intensity, size, and shape. The cell-based regions of interest (ROIs) were then eroded (to reduce the influence of the neuropil signal around the cell bodies as far as possible), and all pixels within each ROI were averaged to yield a time course ($\Delta F/F$) for each neuron; to remove slow fluctuations in the signal, this single time course was subsequently high-pass filtered at a 0.02 Hz cut-off frequency. A fast non-negative deconvolution method approximating the maximum a posteriori spike train for each neuron, given the fluorescence observations³⁴, was used to infer spike trains from calcium signals. This method yields spike probabilities (or inferred firing rate) that are linearly related to the number of action potentials per imaging frame³⁵.

***In vitro* whole-cell recording.** After *in vivo* imaging experiments, whole-cell recordings *in vitro* were performed using an approach as described previously^{10,13}. After two-photon calcium imaging *in vivo*, red fluorescent microspheres (Lumafuor) were injected into the imaged region to facilitate identification of the same region in sliced tissue. We then rapidly removed the mouse brain, dissected it in ice-cold artificial cerebrospinal fluid (ACSF) containing 125 mM NaCl, 2.5 mM KCl, 1 mM MgCl₂, 1.25 mM NaH₂PO₄, 2 mM CaCl₂, 26 mM NaHCO₃, 25 mM dextrose; osmolality 315–325 mOsm, bubbled with 95% O₂/5% CO₂, pH 7.4. Visual cortex slices (300 μ m) were cut coronally on a microtome (VT1200S, Leica Biosystems) and were incubated at 34 °C for thirty minutes before being transferred to the recording chamber. We identified the slice containing the imaged region by the red microsphere injection site and the presence of OGB-1 green fluorescence. To

identify the cells' relative locations, a detailed morphological stack of the slice was obtained with a custom-built microscope and a mode-locked Ti:sapphire laser (Vision-S, Coherent) at 830 nm through a $\times 16$ water immersion objective (0.8 numerical aperture, Nikon). Scanning and image acquisition were controlled by custom software written in LabVIEW (National Instruments). To match the neurons recorded *in vitro* with the neurons imaged *in vivo*, three-dimensional image registration of *in vivo* and *in vitro* image stacks was carried out through an affine transformation using custom-written MATLAB software. We performed simultaneous whole-cell recordings, in 28 °C ACSF, from two to six cells in regions identified in the *in vivo* stack. Recordings were made using Multiclamp 700B amplifiers (Axon Instruments) and acquired using custom software running in MATLAB. Recording pipettes were mounted on remote-controlled motorised micromanipulators (MicroStar, Scientifica), and filled with internal solution containing 5 mM KCl, 115 mM K-gluconate, 10 mM K-HEPES, 4 mM MgATP, 0.3 mM NaGTP, 10 mM Na-phosphocreatine, 40 μ M Alexa Fluor 594; osmolality 290–295 mOsm, pH 7.2. Junction potentials were not corrected. The chloride reversal potential was ~ -85.2 mV. Cells were approached under visual guidance using laser-scanning Dodt contrast and two-photon imaging. To test for the presence of synaptic connections, five spikes at 30 Hz were evoked in each cell, repeated 30 to 120 times, while searching for postsynaptic responses. EPSP amplitudes were calculated by averaging the data points within 1 ms around the first peak depolarization value. After connectivity mapping, step currents from -50 pA to 700 pA were injected at 50 pA increments. Pyramidal neurons were identified according to morphology in Alexa Fluor 594 filled image stacks, spike half-width (>1 ms), regular-spiking pattern on current injection, and in the presence of connections, depolarizing postsynaptic potentials.

Connection probabilities were calculated as the number of connections detected over the number of potential connections assayed. Traces in which large stimulation artefacts occurred were excluded from the analysis. For pairs in which a high quality recording was achieved in only one cell (for example, the other cell was too depolarized/unhealthy, or the seal resistance was less than 1 G Ω), connectivity was assayed only in the direction from the unhealthy cell to the healthy cell, given that spikes could be evoked in both cells. Data from these pairs were included in the analysis of connection probability, but not in the analysis of bidirectional or unidirectional pairs.

***In vivo* whole-cell recordings.** Experiments were performed in 9 C57BL/6 mice of both sexes, aged P29–40. Mice were initially anesthetized with a mixture of fentanyl (0.05 mg per kg), midazolam (5.0 mg per kg), and medetomidine (0.5 mg per kg). The skull was exposed and a metal head-plate was attached with dental cement. A small craniotomy (1–2 mm) was carried out over primary visual cortex, based on stereotaxic coordinates. The dura was removed and the cortex was kept moist with cortex buffer solution (125 mM NaCl, 5 mM KCl, 10 mM glucose, 10 mM HEPES, 2 mM MgSO₄, and 2 mM CaCl₂, pH 7.4).

A silver reference electrode was fixed in place using 2% agarose in cortex buffer under the skin of the neck. Recording pipettes were made using thick-walled filamentous borosilicate glass capillaries (G150F-3, Harvard apparatus) using a horizontal (Sutter, P1000) or vertical (Narashige, PC10) puller adjusted to produce a tip size of approximately 1 μ m and resistance of 5–7 M Ω when filled with intracellular solution containing: 120 mM K-gluconate, 4 mM NaCl, 40 mM HEPES, 2 mM MgATP, 0.3 mM NaGTP; osmolality 295 mOsm, pH 7.4 or 135 mM K-gluconate, 10 mM KCl, 10 mM HEPES, 4 mM MgATP, 0.3 mM NaGTP, 10 mM Na-phosphocreatine; osmolality 295 mOsm, pH 7.4.

Recording pipettes were mounted on a remote-controlled motorised micromanipulator ('Junior', Luigs & Neuman) and orientated at an elevation of approximately 45° from horizontal. The signal was amplified using a Multiclamp 700B amplifier (Molecular Devices), processed by a 50/60 Hz noise eliminator (Humbug, Digitimer) and digitized at 20 kHz by an 18-bit ADC/DAC board (National Instruments) and low-pass filtered at 6 kHz. Data acquisition was controlled by a computer running either Igor Pro (Wavemetrics)/Neuromatic (Jason Rothman, UCL) or a custom MATLAB program. Data were digitally stored for off-line analysis.

Recordings were made using the blind whole-cell patch technique³⁶. Neurons were targeted at a depth of 150–300 μ m below the pial surface (estimated using the reading of the micromanipulator). After a neuron was encountered and whole-cell access was achieved, the recording was switched to current-clamp configuration (0 pA holding current) and data acquired at 10 kHz. Junction potentials were not corrected.

Visual stimuli were generated using MATLAB Psychophysics Toolbox, and displayed on a 43×23 cm LCD monitor with a refresh rate of 60 Hz. The monitor was located 16 cm from the eye, and covered $\sim 106 \times 71$ degrees of the visual space. At the beginning of each experiment, a coarse retinotopic map was acquired for each neuron. Patches of moving black and white gratings were presented in 28 different locations on a grey background, in a pseudorandom order (stimulus duration:

600 ms; interstimulus interval: 400 ms). The monitor position was adjusted to centre the RF on the middle of the screen.

Test visual stimuli consisted of square-wave gratings of 4 orientations drifting in 8 evenly spaced directions spanning 0–360°. Stimulus orientation was perpendicular to drift direction. Stimuli of 50% contrast were presented in a fixed sequence, with the direction of drift increasing in 135° increments. Spatial and temporal frequencies of all stimuli were 0.04 cycles per degree and 3 cycles per s, respectively. Drifting gratings (1.67 s) were interspersed with stationary gratings of 1.67 s. Stationary gratings were of the same orientation as the subsequent drifting grating. The entire stimulus sequence was repeated 5–10 times.

Analysis was restricted to a window commencing 250 ms after the onset of each drifting stimulus and lasting for 1,333 ms (that is, 4 cycles of the 3 Hz square wave grating). The onset of the analysis window was delayed relative to the start of the stimulus in order to avoid the neuron's initial onset response.

Spike thresholds were localized to the maximum second differential within the 5 ms period preceding the action potential peak and spike timing was recorded at 0.1 ms precision. For each spike the subthreshold trace was capped at the threshold value (from the point of threshold until the point at which the membrane potential passed back below this value). If spikes occurred within 10 ms of each other the trace was capped until the membrane potential passed below threshold following the last spike. After spike removal subthreshold traces were smoothed using a 7-ms sliding window.

The baseline membrane potential of each recording block was defined as follows. The pre- and post-stimulus baseline conditions were divided into 1-s sections and the minimum membrane potential observed during each section was averaged to give the baseline of that recording session. The amplitude of depolarization (F0 component) and modulation (F1 component) in response to drifting gratings was measured in the following way: a sine wave function given by $\sigma(t) = A \sin(2\pi ft + \phi) + B$, where f is the temporal frequency of the drifting gratings (3 Hz), and A , B and ϕ are free parameters, was fit to the membrane potential during the analysis window, averaged over all trials of a given stimulus orientation. The amplitude component (F0) was taken as the depolarization B from the baseline membrane potential, and the modulation component (F1) was taken as the amplitude of the sine-wave fit, A .

Receptive field measurement. We displayed natural image sequences (1800 individual images) at 1.4 s intervals (0.4 s presentation time, interleaved by 1 s grey screen). After the onset of each natural image, we recorded 10 imaging frames at ~7 Hz before the next image was presented. For each imaged cell, spike probabilities were inferred from calcium signals using the fast non-negative deconvolution method described above. The response to an image was calculated in the following way: for each visual stimulus, $k(1, \dots, N)$, and each cell, $i(1, \dots, C)$, the response to the stimulus can be expressed $r(k, i, j)$ where $j = 1, \dots, 10$ are the 10 imaging frames. A response value of cell i to stimulus k was then defined as $\frac{\sum_{j=1}^{10} r(k, i, j)}{3}$.

To estimate linear RFs, we used a regularized pseudoinverse method³⁷ to reverse-correlate neuronal responses with natural images. This algorithm regularizes the inverse problem by introducing a two-dimensional smoothness constraint on the linear RF; specifically, the constraint is that the Laplacian of the RF should be close to zero at all points ($\nabla^2 RF = 0$). This method introduces a regularization parameter, λ , which balances the emphasis to be placed on fitting the data and the emphasis to be placed on the smoothness constraint.

The following analysis was performed to choose the regularization parameter. For each cell and each value of the regularization parameter, we separated the natural images and associated responses into training data sets (75% of the data) and test data sets (the remaining 25% of the data). The images included in the training set were chosen randomly and the remaining 25% of the images were placed into the test set. We then calculated linear RFs using the training data, and fit a sigmoid nonlinearity, which can be described by the equation

$$P(x) = \frac{A}{1 + \exp(-\alpha x + \beta)}$$

(where A is the amplitude, α determines the slope, and β determines the offset of the sigmoid) to the training data in order to convert the linear predictions made by the RF into neuronal spike probabilities. We then used the linear RF and nonlinearity to predict responses to the natural images of the test set and took the correlation coefficient between the actual and predicted responses as a measure of RF prediction performance. We performed this procedure 100 times for each cell and each value of the regularization parameter. For each cell, we chose the regularization parameter that maximized the RF prediction performance. Using this procedure, 42% of neurons, 1,969/4,743, had fraction of explained variance >10% (range: 22–59%, 17 experiments), comparable to previous reports (48%, 222/463, in ref. 13), where fraction of explained variance is defined as 100 (R^2), where R is the Pearson's product-moment correlation coefficient between the actual and predicted responses.

To assess whether the RF for a particular cell was significant, we randomly shuffled the response vector to the natural image sequence and performed the reverse correlation again using the same regularization parameter, λ . This procedure was repeated 100 times to produce 100 shuffled RFs, RF_{shuffled} . From these shuffled RFs the mean, μ_{shuffled} and standard deviation, σ_{shuffled} across all pixels were calculated. A RF was defined to be significant if there were pixels which had absolute values $\mu_{\text{shuffled}} + L\sigma_{\text{shuffled}}$, where L (the 'significance level') denotes the number of standard deviations from the mean. The fraction of neurons with RFs with significance level >5 was 53% (2,532/4,743, range: 34–68%, 17 experiments).

The RF was parameterized by fitting a two-dimensional Gabor function using the Levenberg–Marquardt algorithm. The Gabor function is described by

$$G(x', y') = A \exp\left(-\frac{x'^2}{2\sigma_x^2} - \frac{y'^2}{2\sigma_y^2}\right) \cos(2\pi f x' + \phi)$$

where

$$x' = (x - c_x) \cos \theta - (y - c_y) \sin \theta$$

$$y' = (x - c_x) \sin \theta + (y - c_y) \cos \theta$$

These equations describe an underlying two-dimensional cosine grating parameterized by θ (orientation), f (spatial frequency) and ϕ (phase), which is enveloped by a two-dimensional Gaussian function parameterized by A (amplitude), (c_x, c_y) (centre of the Gaussian) and σ_x and σ_y (standard deviations of the Gaussian perpendicular to and parallel to the axis of the grating, respectively).

The pixel–pixel Pearson's correlation coefficient was used as a measure of RF similarity. To compare the difference in orientation preference, the two-dimensional Fourier transform of each RF was taken, $\mathcal{F}\{RF\}$. An orientation tuning curve for each neuron was calculated by interpolating the power value of the Fourier spectrum, $|\mathcal{F}\{RF\}|^2$ around a circle with radius equal to the dominant spatial frequency of the RF (also derived from the two-dimensional Fourier transform). The Gabor fits were used to compare the amount of ON and OFF subfield overlap between pairs of neurons (Fig. 2 and Extended Data Figs 4 and 5). In this case, ON subfields were defined as the region in which pixels of the Gabor fit were >20% of maximum absolute value, $\max(\text{abs}(\text{Gabor fit}))$. Similarly, OFF subfields were defined as the region in which pixels of the Gabor fit were <20% of the negative of the maximum absolute value, $-\max(\text{abs}(\text{Gabor fit}))$. The amount of overlap was defined as

$$\text{overlap} = \frac{|A \cap B|}{|A \cup B|}$$

where A and B are the regions of visual space covered by the presynaptic (say, A) and postsynaptic (say, B) ON, OFF, or both subfields.

General statistical analyses. All statistical tests used in the manuscript were non-parametric, with no assumptions concerning normality or of equal variances. Statistical tests used are described in the manuscript or in the figure legends.

Statistical analysis of connectivity predictions. To determine which aspects of neuronal responses predict connection strength we performed the following analysis. Each pairwise similarity metric (for instance, difference in orientation preference or RF correlation) was ordered such that higher values indicated higher similarity between two neurons. Our test hypothesis (H1) was that there was a positive (linear) correlation between the similarity metric and observed connection weights. Our null hypothesis (H0) was that connections are drawn randomly from the measured distribution of connection strengths. The prediction performance value of a pairwise similarity metric was defined as the Pearson's product-moment correlation coefficient between the similarity metric and the observed synaptic weights, $PP_O = \text{corr}(\mathbf{p}, \mathbf{o})$, where \mathbf{p} and \mathbf{o} are the vectors of the metric and observed connection weights, respectively. To measure the prediction for difference in orientation preference, we combined connection amplitude data from a previous publication, in which orientation was measured using drifting oriented gratings (see refs. 10 and 15), and the RF data, where orientation was measured using the Fourier analysis described above. Only neurons with RFs having a significance level >4.3 (see above) were included in the analysis for subfield overlap and orientation preference.

We estimated P values for this hypothesis using a Monte-Carlo analysis. We generated random permutations of the observed connections and connection strengths (50 permutations were generated for each observed connection). For each permutation, we calculated the prediction performance, $PP_R = \text{corr}(\mathbf{o}, \mathbf{r})$ where \mathbf{o} and \mathbf{r} are the vectors of observed and randomly permuted connection weights, respectively. The P values for hypothesis H1 were then estimated as the proportion of prediction performance values, PP_R that were higher than the value PP_O .

Receptive field transformation. To allow us to pool RF data across neurons, we normalized postsynaptic RFs by first defining a template RF, which was a vertical Gabor with 0 degree phase (that is, centred on an ON domain; see Extended Data Fig. 6). A Gabor was fit to the RF of each postsynaptic neuron, and then rotated,

translated and scaled so that the ON subfield was centred on the template's ON subfield. We used the parameters of this transformation to transform the RFs of all simultaneously recorded presynaptic (whether connected or not) and postsynaptic cells.

The RF outlines used for display (Fig. 3, Extended Data Figs 6, 7 and 8) were calculated in the following way. The maximum pixel value obtained by the RF was defined as Peak_{ON} , and the minimum pixel value obtained by the RF was defined as Peak_{OFF} . If $|\text{Peak}_{\text{ON}}| > |\text{Peak}_{\text{OFF}}|$, then a contour (ON subfield outline) was traced at a value $0.2\text{Peak}_{\text{ON}}$. The OFF subfield outline was then defined as the contour at $(0.7 - 0.5(\text{Peak}_{\text{OFF}}/\text{Peak}_{\text{ON}}))\text{Peak}_{\text{OFF}}$. If $|\text{Peak}_{\text{OFF}}| > |\text{Peak}_{\text{ON}}|$, then a contour (OFF subfield outline) was traced at a value $0.2\text{Peak}_{\text{OFF}}$. The ON subfield outline was then defined as the contour at $(0.7 - 0.5(\text{Peak}_{\text{ON}}/\text{Peak}_{\text{OFF}}))\text{Peak}_{\text{ON}}$.

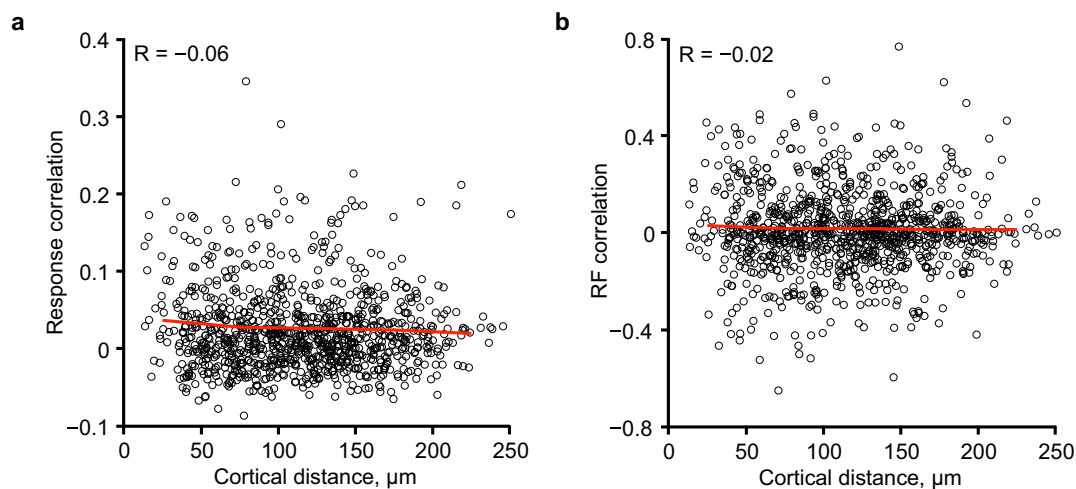
Predicting single-neuron membrane potential responses in a L2/3 network simulation. We used the statistics of recurrent cortical connections measured in our experimental work to predict the feature selectivity of local recurrent inputs to individual L2/3 neurons. The model was used to predict the subthreshold membrane potential of each neuron in response to drifting oriented gratings presented to the network.

To predict inputs in response to drifting gratings, we pooled all RFs measured experimentally into a single population ($n = 4,633$ neurons) by shifting RFs so that the mean RF centre location of each experiment was aligned. Thus, the observed RF scatter for each experiment was retained. We simulated the presentation of 100% contrast, drifting sinusoidal gratings with temporal frequency of 3 Hz (cycles per s) and spatial frequency 0.04 cycles per degree, in 8 equally spaced directions from 0° to 360° .

Network connections were generated by taking each neuron in turn as a postsynaptic neuron, and assigning input connections from all other neurons according to the similarity between the pre- and postsynaptic RFs. We used the relationship between RF correlation and connection amplitude (Fig. 2g) to assign weights to each simulated connection (see schematic in Fig. 4e). Each network simulated the input only to a single postsynaptic neuron. Every network instance was normalized to have the same total weight. The contributions from the strongest 25% of inputs and the weakest 75% of inputs were additionally examined by setting all other weights to zero.

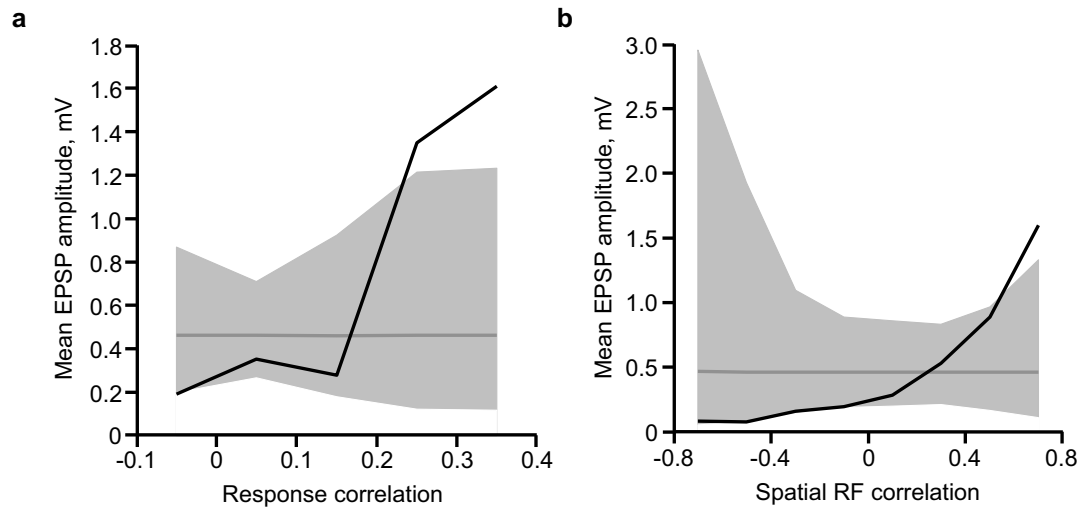
Responses for each neuron in the population were generated using a linear/non-linear model. The parameters of this model were derived for each neuron in the population during estimation of the linear RF (see Methods above). The linear output in response to a particular grating orientation and phase was estimated by correlating the grating stimulus with the RF of a given neuron. The nonlinear sigmoidal transfer curve fit during RF estimation (see Methods above), was used to convert the linear output into a mean firing-rate at each time point of the simulation (Fig. 4e). The firing rates from all presynaptic neurons were then weighted by their connection strengths and summed to generate a time varying membrane potential (V_m) of the postsynaptic neurons (Fig. 4f). No external (for example, feedforward) input was applied to the postsynaptic neuron. The membrane potential for a given postsynaptic neuron was estimated by integrating an R-C membrane model, with input resistance $R_m = 140 \text{ M}\Omega$, capacitance $C = 120 \text{ pF}$. Membrane potential traces produced in this way were analysed identically to membrane potential traces obtained *in vivo* (see above).

30. Mrsic-Flogel, T. D. *et al.* Homeostatic regulation of eye-specific responses in visual cortex during ocular dominance plasticity. *Neuron* **54**, 961–972 (2007).
31. Nimmerjahn, A., Kirchhoff, F., Kerr, J. N. D. & Helmchen, F. Sulforhodamine 101 as a specific marker of astroglia in the neocortex *in vivo*. *Nature Methods* **1**, 1–7 (2004).
32. Brainard, D. H. The psychophysics toolbox. *Spat. Vis.* **10**, 433–436 (1997).
33. Pelli, D. G. The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spat. Vis.* **10**, 437–442 (1997).
34. Vogelstein, J. T. *et al.* Fast nonnegative deconvolution for spike train inference from population calcium imaging. *J. Neurophysiol.* **104**, 3691–3704 (2010).
35. Hofer, S. B. *et al.* Differential connectivity and response dynamics of excitatory and inhibitory neurons in visual cortex. *Nature Neurosci.* **14**, 1045–1052 (2011).
36. Margrie, T. W., Brecht, M. & Sakmann, B. In vivo, low-resistance, whole-cell recordings from neurons in the anaesthetized and awake mammalian brain. *Pflügers Arch.* **444**, 491–498 (2002).
37. Smyth, D., Willmore, B., Baker, G. E., Thompson, I. D. & Tolhurst, D. J. The receptive-field organization of simple cells in primary visual cortex of ferrets under natural scene stimulation. *J. Neurosci.* **23**, 4746–4759 (2003).



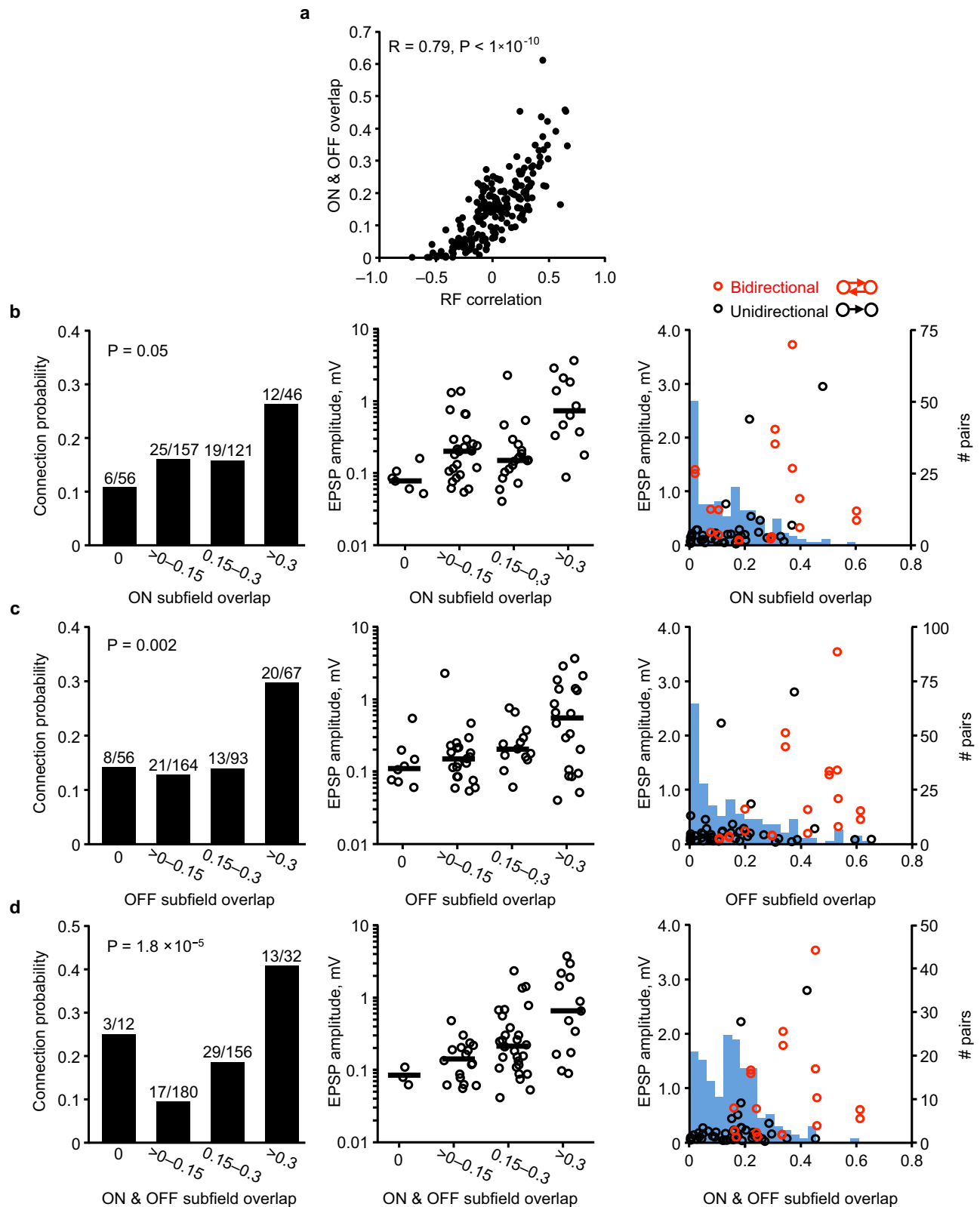
Extended Data Figure 1 | Relationship between response correlation coefficient or RF correlation and cortical distance. **a**, Pairwise response correlation coefficient plotted as a function of cortical distance, for an example region, indicates only a weak relationship between response correlation and

cortical distance ($R = -0.06$). Red line denotes mean value of response correlation in 50 μm bins of cortical distance. **b**, Pairwise RF correlation plotted as a function of cortical distance, for the same example region as in **a**. Again, only a weak relationship was observed ($R = -0.02$).



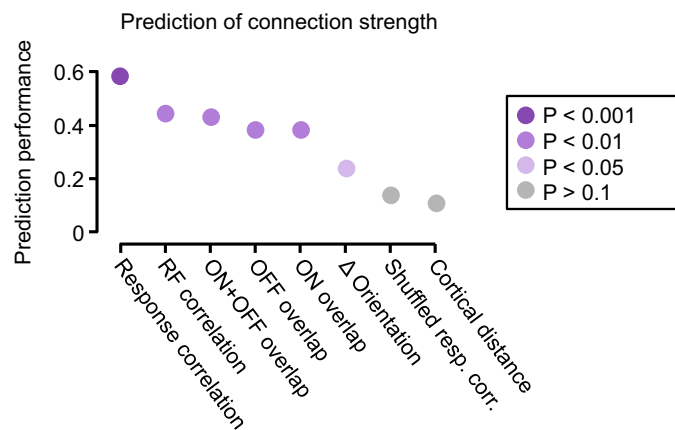
Extended Data Figure 2 | Relationship between mean connection amplitude and response correlation or RF correlation. **a**, Black trace, mean connection amplitude (excluding unconnected pairs) plotted against response correlation. Dashed grey line indicates mean EPSP amplitude of all connections. Grey shaded region represents the 95% confidence interval of the expected mean, estimated by repeated random reshuffling of the EPSP amplitudes among all cell pairs in the data set. Connections were binned with

ranges from -0.1 to 0 , 0 to 0.1 , and so on. **b**, Black trace, mean connection amplitude (excluding unconnected pairs) plotted against RF correlation. Dashed grey line indicates mean EPSP amplitude of all connections. Grey shaded region represents the 95% confidence interval of the expected mean, estimated by repeated random reshuffling of the EPSP amplitudes among all cell pairs in the data set. Connections were binned with ranges from -0.8 to -0.6 , -0.6 to -0.4 , and so on.

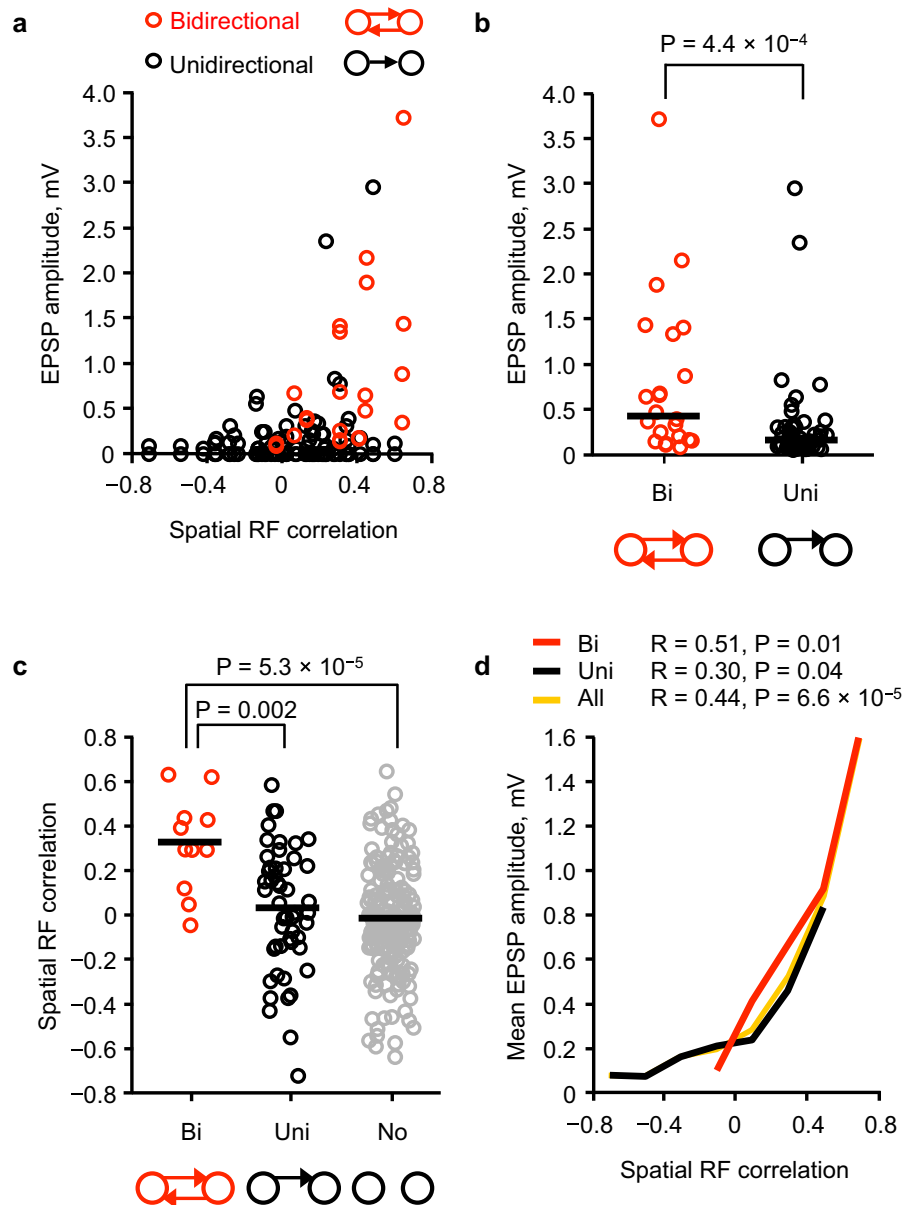


Extended Data Figure 3 | Relationship between connectivity and RF subfield overlap. **a**, The amount of ON and OFF subfield overlap (see Methods) was strongly correlated to the overall RF similarity as measured by RF correlation ($R = 0.79, P < 1 \times 10^{-10}$). **b**, Left panel, connection probability increased with increasing ON subfield overlap ($P = 0.05$; Cochran–Armitage test). Middle panel, EPSP amplitudes categorized into bins of ON overlap. Black line, median EPSP amplitude for each bin. Right panel, EPSP amplitude plotted

against ON overlap. Red data points, bidirectional connections. Black data points, unidirectional connections. Underlying histogram shows frequency of recorded cell pairs as a function of ON overlap. **c**, Same as **b**, but for OFF overlap ($P = 0.002$; Cochran–Armitage test). **d**, Same as **b**, but for combined ON and OFF overlap ($P = 1.8 \times 10^{-5}$; Cochran–Armitage test). P values from the Cochran–Armitage test. To perform the Cochran–Armitage test, the bins at 0 and $>0-0.15$ were considered together, so that groups were evenly spaced.

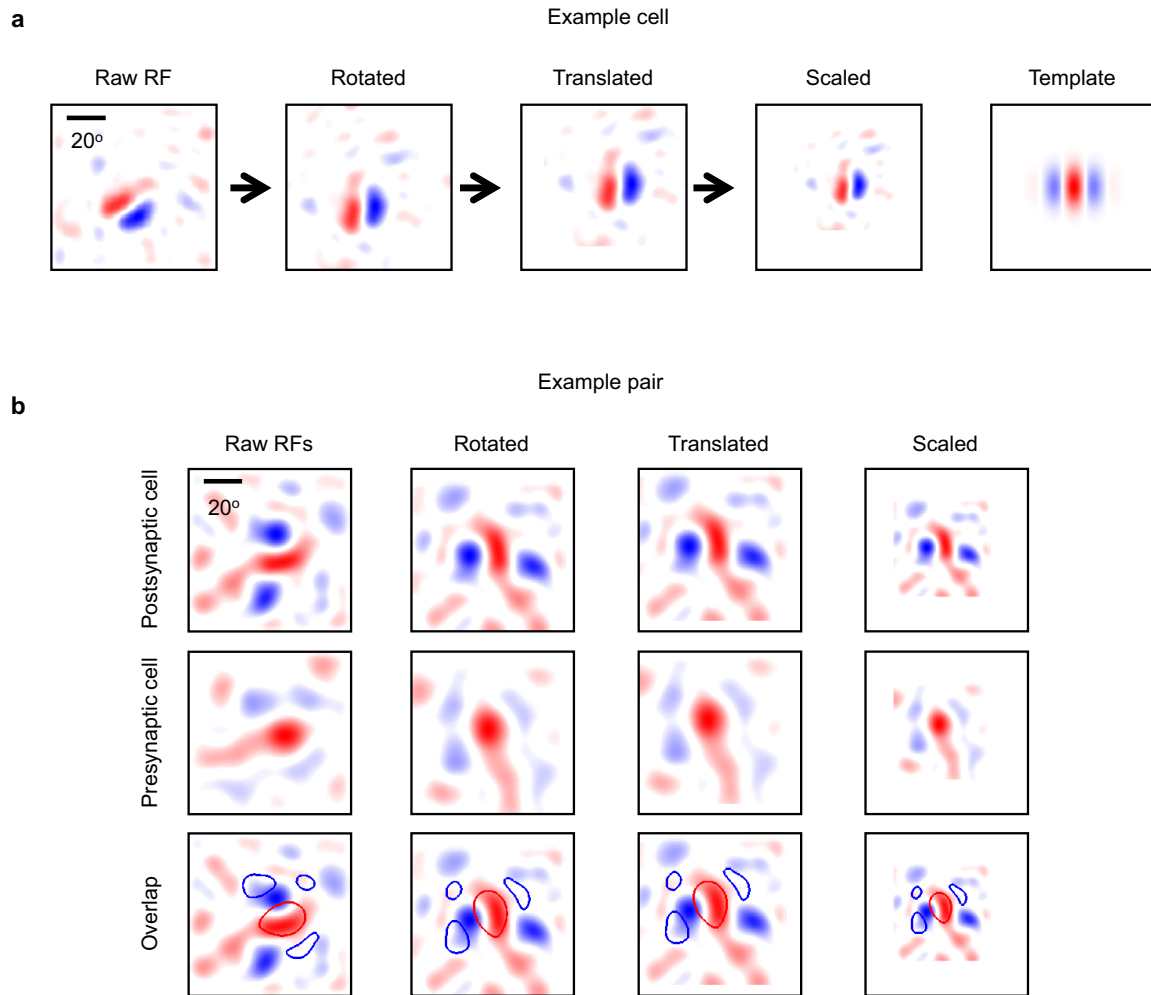


Extended Data Figure 4 | Similarity of shared neuronal properties ranked according to how well they predict connection amplitude, when excluding unconnected pairs. Prediction performance and P values were calculated using a Monte-Carlo analysis (see Methods). Colours of the discs indicate P values.



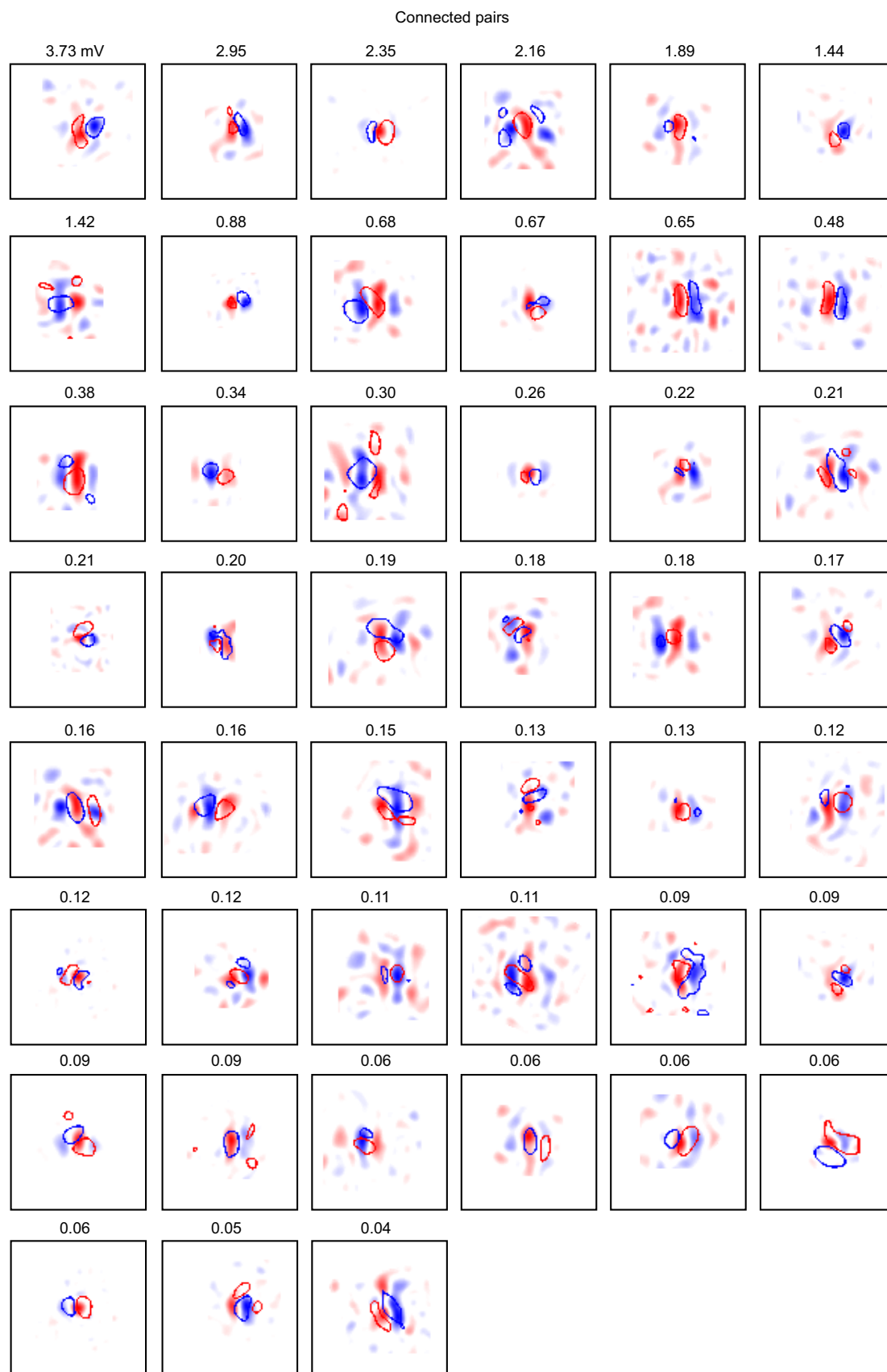
Extended Data Figure 5 | Relationship between bidirectional and unidirectional connections and RF properties. **a**, EPSP amplitude plotted against RF correlation from bidirectionally (red) and unidirectionally connected pairs (black). Replotted from Fig. 2g. **b**, EPSP amplitude for bi- or unidirectional connections. Bidirectional connections were stronger than unidirectional connections (median connection amplitude: 0.44 mV for bidirectional connections, $n = 22$; 0.16 mV for unidirectional connections, $n = 50$; $P = 4.4 \times 10^{-4}$, Wilcoxon rank-sum test). **c**, RF correlation for bidirectionally connected, unidirectionally connected and unconnected pairs. The RFs of bidirectionally connected pairs were more correlated than those of

unidirectionally connected or unconnected pairs (median RF correlation: 0.3 for bidirectionally connected pairs, $n = 11$; 0.04 for unidirectionally connected pairs, $n = 50$; $P = 0.002$; and -0.02 for unconnected pairs, $n = 191$, $P = 5.3 \times 10^{-5}$), although unidirectionally connected pairs did not have higher RF correlations than unconnected pairs ($P = 0.18$, Wilcoxon rank-sum test). **d**, Mean EPSP amplitude versus RF correlation for all (yellow), unidirectionally (black) or bidirectionally (red) connected pairs. There was a positive relationship between RF correlation and connection amplitude for both unidirectional and bidirectional connections.



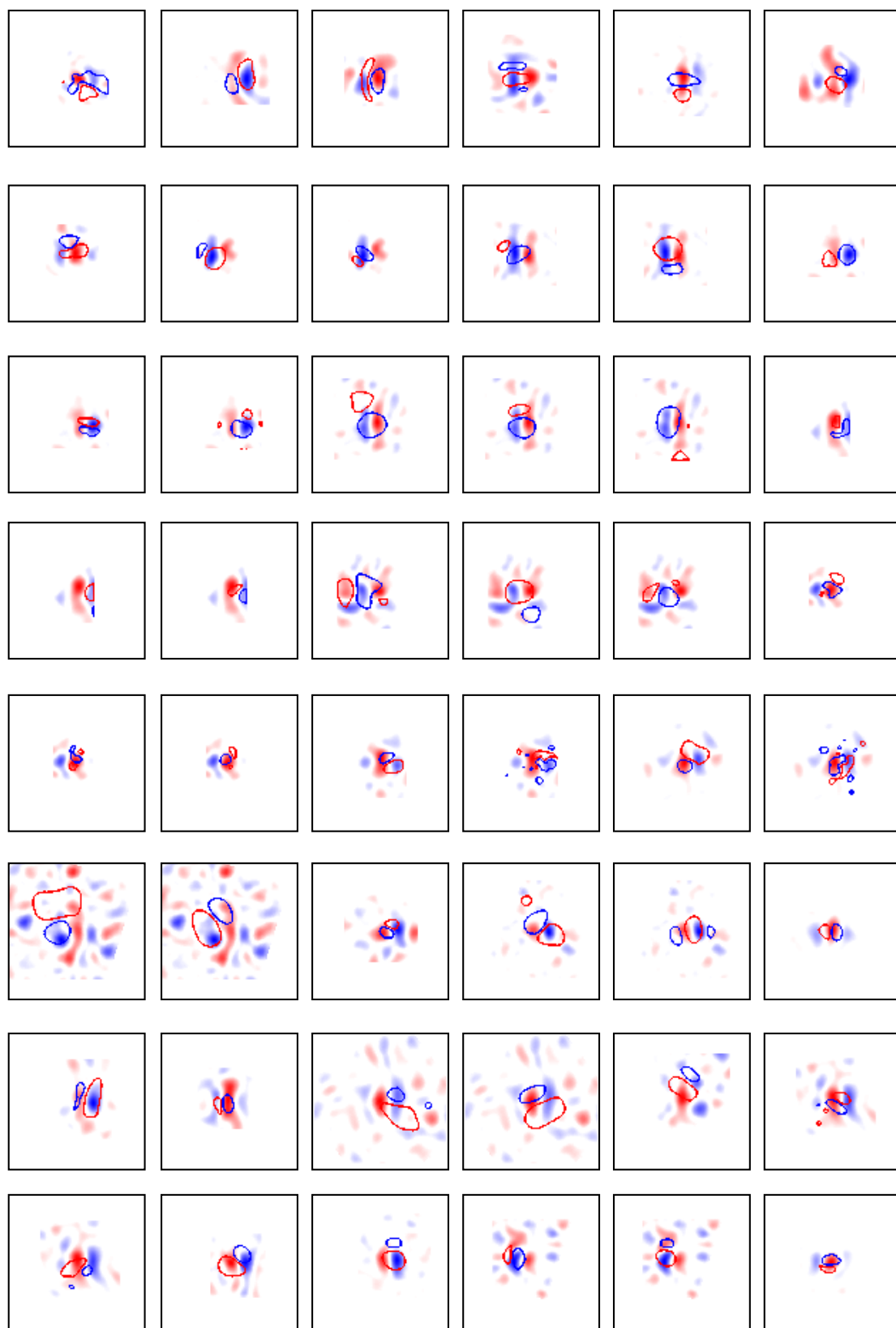
Extended Data Figure 6 | Method of RF normalization. **a**, We normalized postsynaptic RFs to a template RF that was a vertical Gabor with 0 degree phase and an arbitrary but fixed spatial frequency (far right). A Gabor was fit to the RF of each postsynaptic neuron, and then rotated, translated and scaled so that the ON subfield was centred on the template's ON subfield and the spatial

frequencies matched. The same transformation was applied to presynaptic RFs of any simultaneously patched neurons. **b**, Transformation of the RF from an example postsynaptic neuron (upper row), and for the RF for its connected presynaptic neuron (middle row). Bottom row shows presynaptic RF outline overlaid on the postsynaptic RF at each step in the transformation.

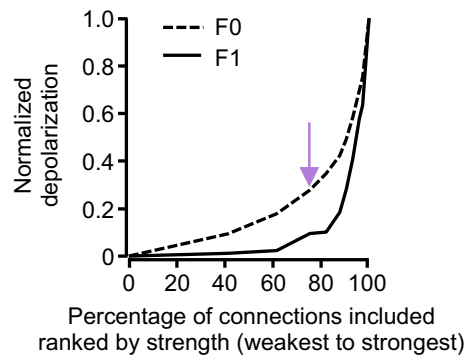


Extended Data Figure 7 | Overlay of RFs between connected neurons. Presynaptic RF outline overlaid on the postsynaptic RF for all the connected pairs after performing normalization of the pre- and postsynaptic RFs to the RF template ($n = 45$). Numbers indicate connection amplitude.

Unconnected pairs



Extended Data Figure 8 | Overlay of RFs between unconnected neurons. Assessed presynaptic RF outlines overlaid on the assessed postsynaptic RF for a representative set of unconnected pairs after normalization to the RF template.



Extended Data Figure 9 | Contribution of strong and weak connections to membrane potential depolarization. Removal of an increasingly larger fraction of the strongest inputs from the L2/3 model steeply reduces the large modulation component (F1) but more gradually reduces the mean depolarization component (F0). Model from Fig. 4d. Purple arrow indicates the weakest 75% of connections, as shown in Fig. 4i, j.

Modulation of the proteoglycan receptor PTP σ promotes recovery after spinal cord injury

Bradley T. Lang¹, Jared M. Cregg¹, Marc A. DePaul¹, Amanda P. Tran¹, Kui Xu², Scott M. Dyck³, Kathryn M. Madalena¹, Benjamin P. Brown⁴, Yi-Lan Weng⁵, Shuxin Li⁶, Soheila Karimi-Abdolrezaee³, Sarah A. Busch¹, Yingjie Shen² & Jerry Silver¹

Contusive spinal cord injury leads to a variety of disabilities owing to limited neuronal regeneration and functional plasticity. It is well established that an upregulation of glial-derived chondroitin sulphate proteoglycans (CSPGs) within the glial scar and perineuronal net creates a barrier to axonal regrowth and sprouting^{1–5}. Protein tyrosine phosphatase σ (PTP σ), along with its sister phosphatase leukocyte common antigen-related (LAR) and the nogo receptors 1 and 3 (NgR), have recently been identified as receptors for the inhibitory glycosylated side chains of CSPGs^{6–8}. Here we find in rats that PTP σ has a critical role in converting growth cones into a dystrophic state by tightly stabilizing them within CSPG-rich substrates. We generated a membrane-permeable peptide mimetic of the PTP σ wedge domain that binds to PTP σ and relieves CSPG-mediated inhibition. Systemic delivery of this peptide over weeks restored substantial

serotonergic innervation to the spinal cord below the level of injury and facilitated functional recovery of both locomotor and urinary systems. Our results add a new layer of understanding to the critical role of PTP σ in mediating the growth-inhibited state of neurons due to CSPGs within the injured adult spinal cord.

Our laboratory has developed an *in vitro* model of the inhibitory extracellular matrix that forms after spinal cord injury (SCI), wherein adult sensory neurons form dystrophic endballs as they attempt to traverse an increasing gradient of CSPG⁹. Our previous studies focused on dystrophic growth cones stalled within the CSPG gradient that remained active as they recycled membrane^{9,10}. We now report that chronic exposure to CSPG can induce the development of a distinct over-adhered morphology with no forward motility (Fig. 1a–c and Supplementary Videos 1–3). Any newly formed growth cones rapidly involute into large

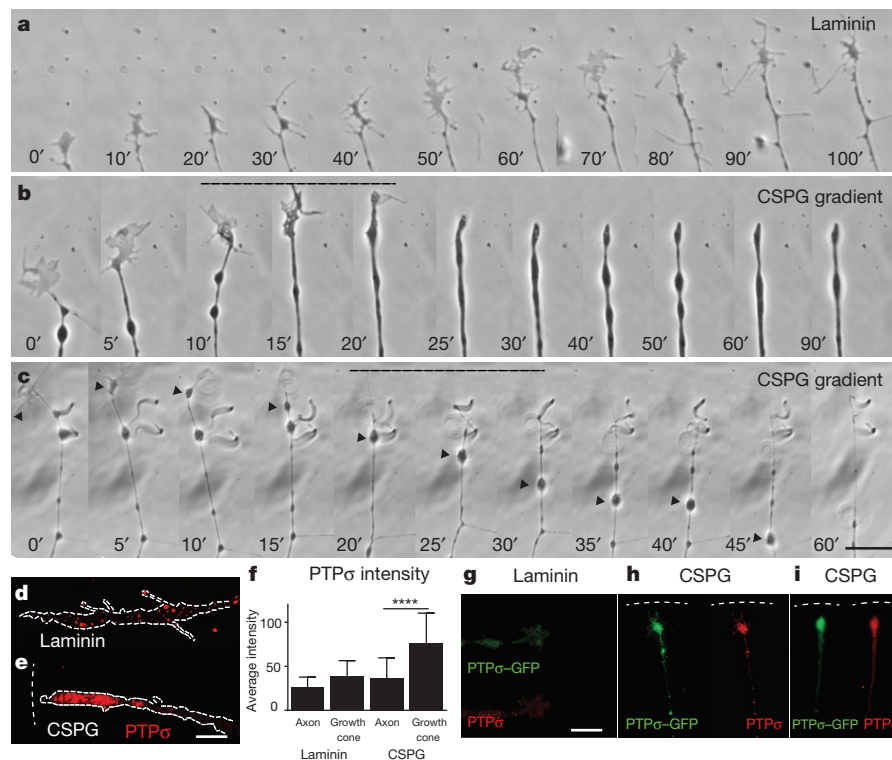


Figure 1 | Immobilization of growth cones within gradients of CSPG. a–c, Time-lapse imaging in which the growth cones of adult sensory neurons are either motile upon a uniform substrate of laminin (a) or stabilizing within the gradient of proteoglycan (b, c). Arrowhead indicates absorbed growth cone. Timestamp indicates time in minutes. Scale bar, 20 μ m. See also Supplementary Videos 1–3. d–j, PTP σ staining in motile or stabilized growth

cones. $n = 16$ laminin, $n = 26$ CSPG for both treatments. Error bars show standard error of the mean (s.e.m.). **** $P < 0.001$, $F = 19.9$, one-way analysis of variance (ANOVA), Tukey's post hoc test. g–i, Adult sensory neurons expressing a PTP σ -green fluorescent protein (GFP) plasmid. Dashed line indicates CSPG gradient.

¹Department of Neurosciences, Case Western Reserve University School of Medicine, Cleveland, Ohio 44106, USA. ²Center for Brain and Spinal Cord Repair, Department of Neuroscience, Wexner Medical Center at The Ohio State University, Columbus, Ohio 43210, USA. ³Regenerative Medicine Program and Department of Physiology, University of Manitoba, Winnipeg, Manitoba R3E 0J9, Canada. ⁴Baldwin Wallace University, Berea, Ohio 44017, USA. ⁵Institute for Cell Engineering, Johns Hopkins University School of Medicine, Baltimore, Maryland, 21205, USA. ⁶Shriners Hospital's Pediatric Research Center (Center for Neural Repair and Rehabilitation), Temple University School of Medicine, Philadelphia, Pennsylvania 19140, USA.

blebs, resulting in a beaded axon with one tip or multiple side branches, all terminating in small punctate adhesive contacts (Fig. 1c and Supplementary Video 3). This morphology is remarkably similar to that described in the early 20th century by Ramon y Cajal in the chronically injured cat spinal cord¹¹.

Interestingly, although PTP σ is evenly distributed in a punctate pattern within motile axons and growth cones, it becomes concentrated in dystrophic stabilized growth cones (Fig. 1d–i). We found similar elevations of LAR, but not NgRs (Extended Data Fig. 1a, b). In addition, we observed a large concentration of PTP σ in the lesion penumbra after SCI (Extended Data Fig. 1c, d). As PTP σ co-localizes with adhesion plaques and focal adhesions^{12,13}, we hypothesized that it had a critical role in growth cone immobilization and progression into a dystrophic state. Therefore, we sought to target PTP σ to relieve CSPG-mediated inhibition.

Upon analysing the structure of PTP σ and related phosphatases, we identified a highly conserved 24-amino-acid intracellular wedge domain (Fig. 2a and Extended Data Fig. 2a, b). As wedge domains are known to regulate downstream signalling through a variety of mechanisms^{7,14–16}, we designed intracellular sigma peptide (ISP), a novel peptide-mimetic of the PTP σ wedge with a TAT domain to facilitate membrane penetration (Fig. 2b).

ISP was able to bind to recombinant human PTP σ (Fig. 2c). In rodent brain and spinal cord lysates, ISP pulled down both immature full-length PTP σ and the mature functional complex (Fig. 2d–f)¹². In PTP σ -null mice, only a very minor signal was detected, which may reflect non-specific binding to PTP δ , the third LAR family member (Fig. 2d, e)¹⁷. No detectable binding was observed between ISP and other CSPG receptors such as LAR and NgRs (Extended Data Fig. 2e, f). Interestingly, a LAR wedge-domain peptide (ILP)¹⁴ was also capable of binding PTP σ , but less efficiently than ISP (Fig. 2d, f and Extended Data Fig. 2c, d).

We next asked whether ISP could release CSPG-mediated axonal inhibition *in vitro*. ISP treatment allowed adult sensory neurons to extend axons through a CSPG gradient in a dose-dependent manner to the same extent as pre-treatment with chondroitinase ABC (ChABC), which cleaves the glycosylated CSPG side chains and removes the PTP σ ligand (Fig. 2g–i)⁶. Time-lapse microscopy revealed that although growth cones treated with ISP were still transiently collapsed by CSPG, they continued to reform growth cones, allowing them to eventually cross the gradient (Fig. 2k and Supplementary Video 4). Additionally, both ISP and ChABC treatments were sufficient to convert already stabilized dystrophic growth cones into a motile state (data not shown). Interestingly, human ISP and the wedge domain peptides of PTP δ and LAR (IDP and ILP) demonstrated some efficacy, suggesting functional redundancy among LAR

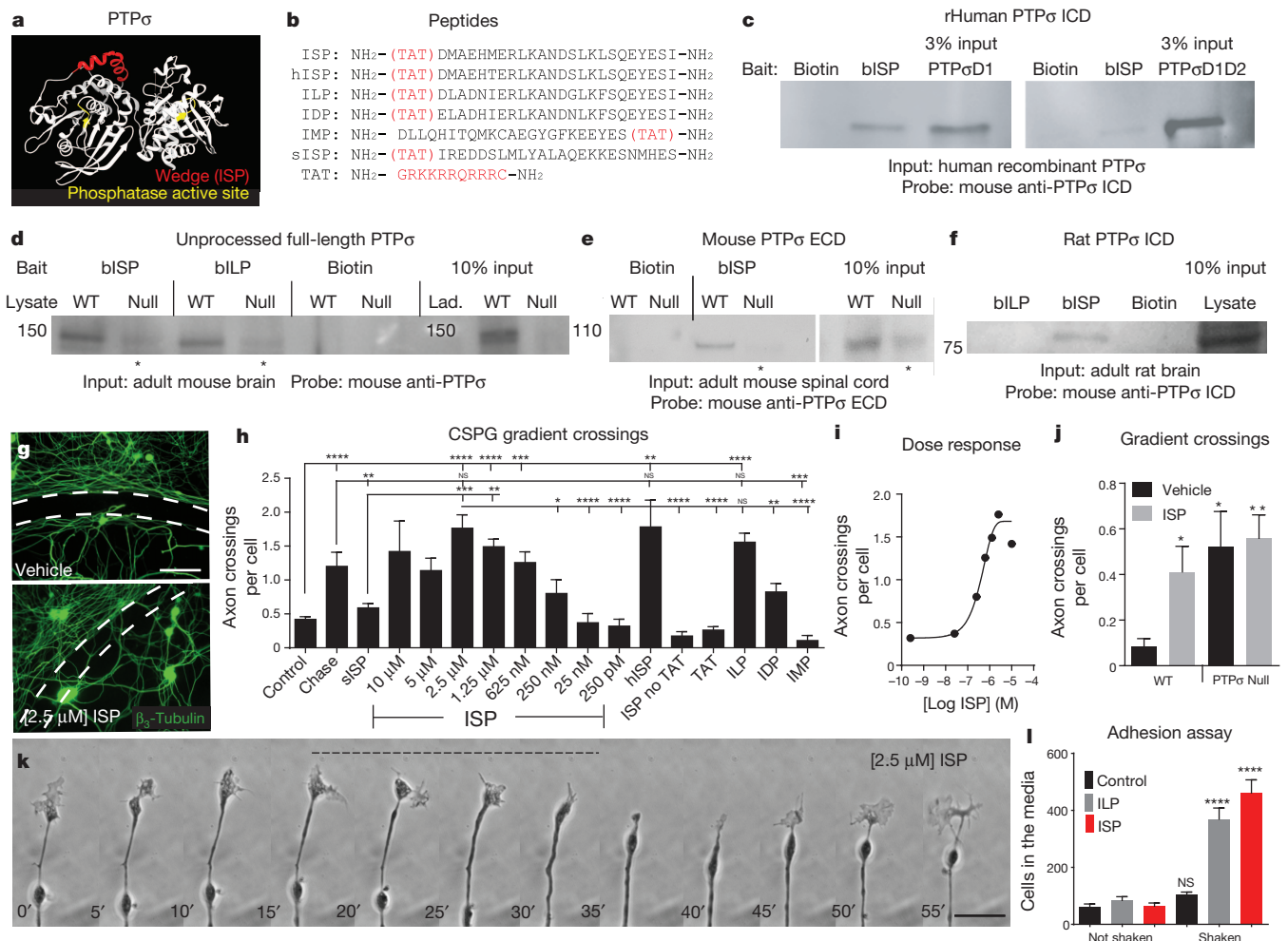


Figure 2 | Identification and characterization of ISP. **a**, PTP σ structure and wedge domain (red). **b**, Peptide sequences. **c–f**, Pull-down of human, rat and mouse PTP σ with biotinylated ISP (bISP). Asterisk indicates nonspecific recognition of PTP δ . ICD, intracellular domain; D1, PTP σ domain 1; D1D2, PTP σ domain 1–2; Lad, ladder (kilodaltons); r, recombinant; WT, wild type. Four repetitions. **g–i**, CSPG gradient crossing assay. Dashed lines indicate CSPG gradient. Scale bar, 50 μ m; $n > 16$ gradients per group. hISP, human ISP; sISP, scrambled ISP. **j**, ISP treatment on PTP σ -null neurons

($n = 12$ per group, 3 repetitions). **k**, Time-lapse imaging of an adult sensory neuron growth cone after 2.5 μ M ISP treatment (Supplementary Video 4). Timestamp indicates time in minutes. Scale bar, 20 μ m. **l**, The number of neurons released from a CSPG-rich substrate after agitation ($n = 28$ vehicle/ILP, $n = 16$ ISP wells per group). Scale bar, 50 μ m. Error bars show s.e.m. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, **** $P < 0.0001$, one-way ANOVA, Tukey's post hoc test. For additional sample size information, see Methods.

family wedge domains. The wedge domain of PTP μ (IMP), scrambled ISP, TAT alone, or ISP without TAT were ineffective (Fig. 2h).

Importantly, treatment of adult PTP σ -null sensory neurons with ISP did not further increase crossings (Fig. 2j)⁶. In addition, mature astrocytes, which have immunocytochemically undetectable PTP σ ¹⁸, were unable to traverse the CSPG gradient after ISP treatment (Extended Data Fig. 3a). However, PTP σ -expressing satellite glia were induced to cross successfully (Extended Data Fig. 3b). These results suggest specificity of ISP interactions through PTP σ .

ISP concentrations above the optimal dose for gradient crossing greatly reduced neuronal cell attachment, suggesting a reversal of the over-adhered phenotype. ISP treatment decreased interactions between the CSPG-rich substrate and PTP σ , allowing cells to detach in response to agitation (Fig. 2l and Extended Data Fig. 3c, d). This suggests that appropriate levels of adhesion are critical for regeneration across increasing concentrations of innately inhibitory CSPGs¹⁹.

The extracellular-signal-regulated kinase 1/2 (Erk1/2) cascade regulates a variety of processes, including axonal growth²⁰. Erk1/2 phosphorylation was decreased in neuronal cells grown on a CSPG-rich substrate and both ISP and ChABC were able to restore the phosphorylation

state of Erk1/2 to levels comparable to that on laminin-only substrates (Extended Data Fig. 4)²¹.

TAT-conjugated peptides are known to cross biological membranes, including the blood–brain barrier²². One hour after a single subcutaneous injection of fluorescein isothiocyanate (FITC)-conjugated ISP, we were able to visualize ISP in the intact nervous system (Extended Data Fig. 5a, b). Therefore, we chose a non-invasive systemic treatment paradigm using daily subcutaneous injections, avoiding complications associated with intraparenchymal delivery²³. Beginning 1 day after contusive SCI, rats were treated with ISP (11 μ g per day), ILP (11 μ g per day) or vehicle once daily for 7 consecutive weeks (Extended Data Fig. 5c–e).

SCI disrupts connections between the bladder and brainstem micturition control centre, leading to reduced void frequency and an accompanying increase in void volume (Extended Data Fig. 6a, b)²⁴. Twelve weeks after injury, ISP promoted a significant twofold increase in void frequency versus controls, along with a significant decrease in void volume (Fig. 3a–c and Extended Data Fig. 6c–h). We used urodynamics to assess whether this improvement was a result of physiologically normal bladder activity. Rats urinate by contracting the detrusor muscle, resulting in a rapid increase in bladder pressure while the external urethral sphincter

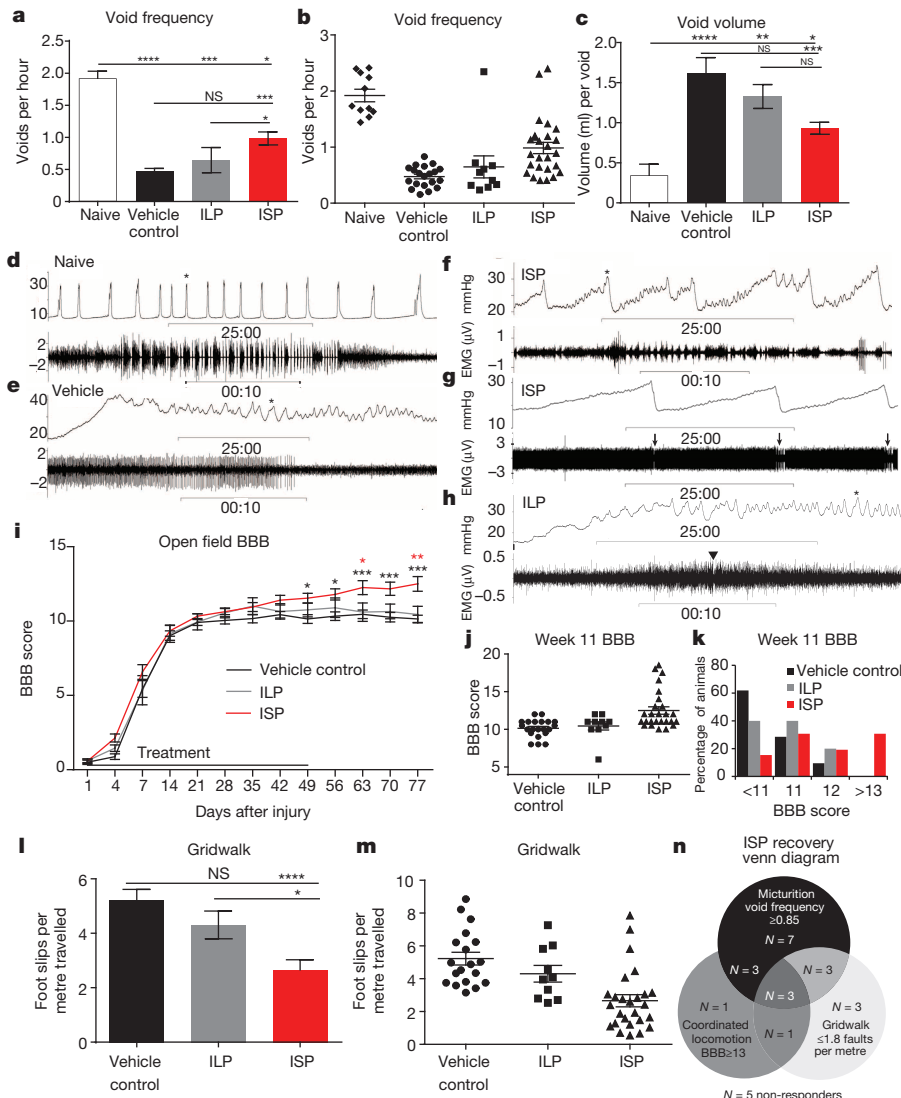


Figure 3 | Functional recovery after ISP treatment. **a–c**, Void frequency and average void volume at 12 weeks after SCI. **d–h**, Representative urodynamic recordings of detrusor activity (bladder pressure, top trace) and EUS activity (bottom trace, expanded at the points marked with an asterisk). **g**, Full trace of EUS activity. Arrows indicate synchronized phasic bursting; arrowhead indicates single burst. **i–k**, Locomotor recovery (BBB score) after

SCI. **l, m**, Gridwalk test at 12 weeks after SCI. **n**, ISP functional recovery Venn diagram (21/26 ISP, 0/21 vehicle, 1/10 ILP). Error bars show s.e.m. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, **** $P < 0.0001$, repeated measures two-way ANOVA, Tukey's post-hoc test (BBB); one-way ANOVA, Kruskal–Wallis post-hoc test (gridwalk and micturition). NS, not significant. Black indicates ISP versus control; red indicates ISP versus ILP.

(EUS) bursts phasically, expelling urine (Fig. 3d)²⁵. No vehicle- or ILP-treated animals displayed multiple coordinated bursts of the EUS during a void (Fig. 3e, h). Although the detrusor was still hyperactive, 10 of the 15 analysed ISP-treated animals recovered coordination between bladder contractions and EUS phasic bursting, suggesting reconnection of a functional circuit (Fig. 3f, g)²⁶.

We measured locomotor recovery using the Basso, Beattie and Bresnahan (BBB) scale and a gridwalk test²⁷. After SCI, vehicle- and ILP-treated animals recovered from hindlimb paralysis at day 1 to, on average, occasional weight bearing stepping at week 11 (Fig. 3i). ISP treatment resulted in a significant progressive recovery of locomotion, which began several weeks after injury (Fig. 3i). Thirty per cent of ISP-treated animals (versus zero vehicle/ILP-treated animals) demonstrated, at least, frequent coordinated stepping (BBB ≥ 13), with three rats achieving BBB scores ≥ 17.5 (Fig. 3j–k). Furthermore, ISP-treated animals made on average 58% fewer foot faults than control rats on the gridwalk test (Fig. 3l, m), suggesting recovery of sensorimotor coordination and balance. Importantly, aside from minor irritation at the injection site after multiple weeks of treatment, ISP did not induce neuropathic pain (Extended Data Fig. 7a, b).

Interestingly, no correlation between ISP-induced recovered behaviours was observed (Extended Data Fig. 7c–e). Using stringent threshold analyses, we determined that 21 of 26 ISP-treated animals recovered at least one behaviour, with 3 animals recovering all three (Fig. 3n). Further analyses demonstrated the degree to which responding animals benefited from treatment (Extended Data Fig. 7f–h). Taken together, this suggests that the re-acquisition of each behaviour is modular and not coincident, and may reflect anatomical differences in the pattern of axon re-innervation.

The behavioural results represent the average of five repetitions, each performed with newly synthesized peptide, blinded behavioural testers and a separate blinded treatment administrator. Although variability

existed, ISP increased functional recovery of each behaviour in all cohorts (Extended Data Fig. 8). In a final group of animals, urinary, but not locomotor behaviours, responded to increasing concentrations of ISP, with our maximum (44 μg per day) dose improving urination markedly in all rats (Extended Data Fig. 7g). Therefore, further optimization of the dose or administrative route of ISP may lead to additional functional improvements.

ISP treatment was not neuroprotective, as it did not lead to differences in lesion size or spared white matter (Fig. 4a, b). At the individual animal level, the variability in spared tissue correlated with functional recovery in vehicle-, but not ISP-treated rats (Extended Data Fig. 9).

Although regenerative pathways are difficult to examine after contusive SCI²⁸, we did not observe any lengthy regenerating biotinylated dextran amine (BDA)-labelled corticospinal tract fibres (data not shown). We focused further analysis on the serotonergic system (5HT) neurons, which express LAR family receptors and contribute to proper neuromodulatory tone in locomotion and micturition circuitry^{7,26,29,30}. Contusive SCI led to a marked decrease in descending 5HT-positive fibres caudal to the lesion (Fig. 4c, e, h). In animals exhibiting functional recovery after ISP treatment, we observed unusually shaped, densely sprouted territories of 5HT fibres below the level of the lesion (Fig. 4d, f, i–n). These patterns corresponded in part with neurofilament staining (Extended Data Fig. 10), but were not present in sections stained for GFAP, ED1 or 4',6-diamidino-2-phenylindole (DAPI), ensuring that the sharp edges of staining were not due to tissue folds. We speculate that physical constraints of the perineuronal net may confine fibres in these patterns. Their spatial variability, in conjunction with differences in tract sparing from animal to animal after contusion, could partially account for the disparity in behavioural recovery between animals.

Treatment with the 5HT receptor antagonist methysergide at 14 weeks after SCI significantly reduced locomotor and urinary function in ISP-treated, but not vehicle-treated animals (Fig. 4o, p)^{26,29}. This was most

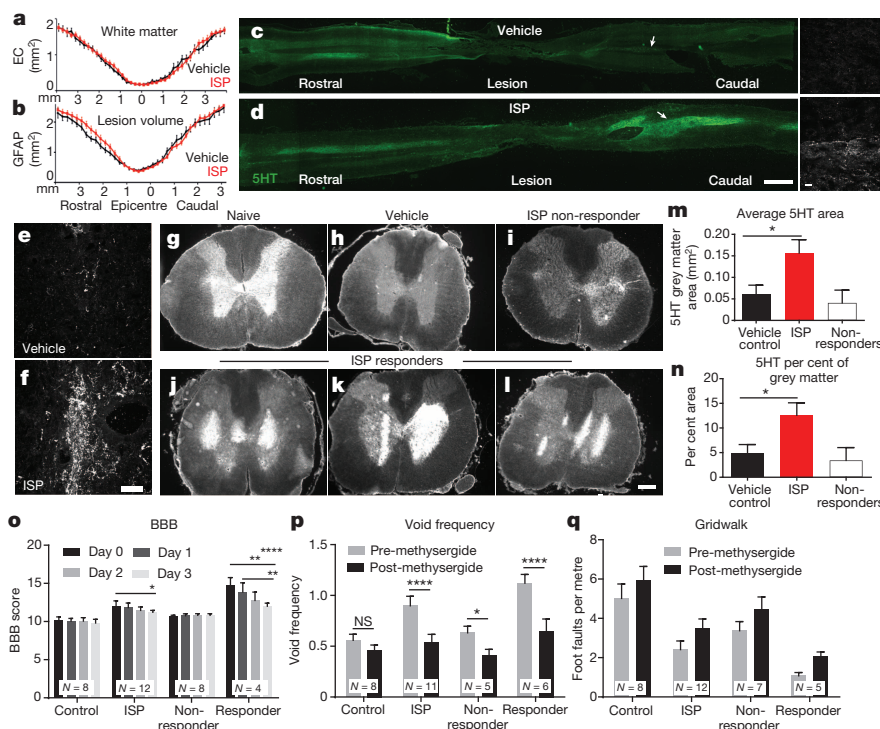


Figure 4 | Anatomical changes after ISP treatment. **a, b**, Average lesion size (GFAP) and spared white matter (eriochrome cyanine (EC) staining) after SCI. $n = 9$ –10 vehicle, $n = 14$ ISP. **c, d**, 5HT intensity in representative longitudinal spinal cord sections. Insets reveal fibre density. Scale bar, 2 mm; inset, 20 μm . **e, f**, Representative confocal projections of caudal 5HT fibres at L1–L3. Scale bar, 20 μm . **g–l**, 5HT intensity across lumbar coronal sections. Scale bar, 500 μm . **m, n**, Average area and per cent coverage of 5HT. $n = 13$

vehicle, $n = 18$ ISP. $P < 0.05$, Student's t -test (two tailed). **o–q**, Behavioural response to methysergide at 14 weeks after SCI. 'Responders' are animals demonstrating functional recovery in each behaviour. $n = 8$ vehicle, $n = 11$ –12 ISP. Error bars show s.e.m. * $P < 0.05$, ** $P < .01$, *** $P < 0.001$, two-way repeated measures ANOVA, Tukey's post-hoc test (BBB); one-way ANOVA, Kruskal–Wallis post hoc test (gridwalk and void frequency). NS, not significant.

evident in ISP responders, animals that regained function beyond recovery thresholds. However, behavioural improvements were not fully reverted to vehicle levels and gridwalk scores were only minimally affected (Fig. 4q), suggesting plasticity of other pathways outside the serotonergic system.

While CSPGs have largely been thought to act as repulsive components of the extracellular matrix, our results suggest that regenerating adult growth cones can become permanently immobilized by CSPG gradients. These observations highlight an initial cellular mechanism regulated by PTP σ that leads to the development of axonal dystrophy and the prevention of chronic regeneration and plasticity *in vivo*. Systemic modulation of PTP σ opens a new therapeutic avenue in non-invasive treatments for enhancing functional recovery after a variety of injuries or diseases in which proteoglycans inhibit the attempt of axons to regenerate or sprout.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 8 July; accepted 16 October 2014.

Published online 3 December 2014.

- Cregg, J. M. *et al.* Functional regeneration beyond the glial scar. *Exp. Neurol.* **253**, 197–207 (2014).
- Andrews, E. M., Richards, R. J., Yin, F. Q., Viapiano, M. S. & Jakeman, L. B. Alterations in chondroitin sulfate proteoglycan expression occur both at and far from the site of spinal contusion injury. *Exp. Neurol.* **235**, 174–187 (2012).
- Pizzorusso, T. *et al.* Reactivation of ocular dominance plasticity in the adult visual cortex. *Science* **298**, 1248–1251 (2002).
- Bradbury, E. J. *et al.* Chondroitinase ABC promotes functional recovery after spinal cord injury. *Nature* **416**, 636–640 (2002).
- Massey, J. M. *et al.* Chondroitinase ABC digestion of the perineuronal net promotes functional collateral sprouting in the cuneate nucleus after cervical spinal cord injury. *J. Neurosci.* **26**, 4406–4414 (2006).
- Shen, Y. *et al.* PTP σ is a receptor for chondroitin sulfate proteoglycan, an inhibitor of neural regeneration. *Science* **326**, 592–596 (2009).
- Fisher, D. *et al.* Leukocyte common antigen-related phosphatase is a functional receptor for chondroitin sulfate proteoglycan axon growth inhibitors. *J. Neurosci.* **31**, 14051–14066 (2011).
- Dickendesher, T. L. *et al.* NgR1 and NgR3 are receptors for chondroitin sulfate proteoglycans. *Nature Neurosci.* **15**, 703–712 (2012).
- Tom, V. J., Steinmetz, M. P., Miller, J. H., Doller, C. M. & Silver, J. Studies on the development and behavior of the dystrophic growth cone, the hallmark of regeneration failure, in an *in vitro* model of the glial scar and after spinal cord injury. *J. Neurosci.* **24**, 6531–6539 (2004).
- Busch, S. A., Horn, K. P., Silver, D. J. & Silver, J. Overcoming macrophage-mediated axonal dieback following CNS injury. *J. Neurosci.* **29**, 9967–9976 (2009).
- Cajal, S. R. Y. *Degeneration & Regeneration of the Nervous System* (Oxford Univ. Press, 1928).
- Aicher, B., Lerch, M. M., Muller, T., Schilling, J. & Ullrich, A. Cellular redistribution of protein tyrosine phosphatases LAR and PTP σ by inducible proteolytic processing. *J. Cell Biol.* **138**, 681–696 (1997).
- Serra-Pagès, C. *et al.* The LAR transmembrane protein tyrosine phosphatase and a coiled-coil LAR-interacting protein co-localize at focal adhesions. *EMBO J.* **14**, 2827–2838 (1995).
- Xie, Y. *et al.* Protein-tyrosine phosphatase (PTP) wedge domain peptides: a novel approach for inhibition of PTP function and augmentation of protein-tyrosine kinase function. *J. Biol. Chem.* **281**, 16482–16492 (2006).
- Jiang, G. *et al.* Dimerization inhibits the activity of receptor-like protein-tyrosine phosphatase- α . *Nature* **401**, 606–610 (1999).
- Barr, A. J. *et al.* Large-scale structural analysis of the classical human protein tyrosine phosphatome. *Cell* **136**, 352–363 (2009).
- Wallace, M. J., Fladd, C., Batt, J. & Rotin, D. The second catalytic domain of protein tyrosine phosphatase δ (PTP δ) binds to and inhibits the first catalytic domain of PTP σ . *Mol. Cell Biol.* **18**, 2608–2616 (1998).
- Silver, D. J. *et al.* Chondroitin sulfate proteoglycans potently inhibit invasion and serve as a central organizer of the brain tumor microenvironment. *J. Neurosci.* **33**, 15603–15617 (2013).
- Lowery, L. A. & Van Vactor, D. The trip of the tip: understanding the growth cone machinery. *Nature Rev. Mol. Cell Biol.* **10**, 332–343 (2009).
- Polleux, F. & Snider, W. Initiating and growing an axon. *Cold Spring Harb. Perspect. Biol.* **2**, a001925 (2010).
- Sapieha, P. S. *et al.* Receptor protein tyrosine phosphatase sigma inhibits axon regrowth in the adult injured CNS. *Mol. Cell. Neurosci.* **28**, 625–635 (2005).
- Banks, W. A., Robinson, S. M. & Nath, A. Permeability of the blood–brain barrier to HIV-1 Tat. *Exp. Neurol.* **193**, 218–227 (2005).
- Jones, L. L. & Tuszynski, M. H. Chronic intrathecal infusions after spinal cord injury cause scarring and compression. *Microsc. Res. Tech.* **54**, 317–324 (2001).
- de Groat, W. C. *et al.* Mechanisms underlying the recovery of urinary bladder function following spinal cord injury. *J. Auton. Nerv. Syst.* **30** (suppl.), S71–S77 (1990).
- Pikov, V. & Wrathall, J. R. Coordination of the bladder detrusor and the external urethral sphincter in a rat model of spinal cord injury: effect of injury severity. *J. Neurosci.* **21**, 559–569 (2001).
- Lee, Y. S. *et al.* Nerve regeneration restores supraspinal control of bladder function after complete spinal cord injury. *J. Neurosci.* **33**, 10591–10606 (2013).
- Basso, D. M., Beattie, M. S. & Bresnahan, J. C. A sensitive and reliable locomotor rating scale for open field testing in rats. *J. Neurotrauma* **12**, 1–21 (1995).
- Tuszynski, M. H. & Steward, O. Concepts and methods for the study of axonal regeneration in the CNS. *Neuron* **74**, 777–791 (2012).
- Murray, K. C. *et al.* Recovery of motoneuron and locomotor function after spinal cord injury depends on constitutive activity in 5-HT_{2C} receptors. *Nature Med.* **16**, 694–700 (2010).
- Xu, B. *et al.* Role of CSPG receptor LAR phosphatase in restricting axon regeneration after CNS injury. *Neurobiol. Dis.* **73C**, 36–48 (2014).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported by National Institute of Neurological Disorders and Stroke grant NS025713 (J.S.); the Case Western Reserve University Council to Advance Human Health; P. Jing, R. Senior and S. Poon; Unite to Fight Paralysis; The Brumagin Memorial Fund; Spinal Cord Injury Sucks; United Paralysis Foundation; and The Kaneko Family Fund. The authors thank J. Flanagan, M. Blackmore, A. Filous, S. Brady-Kalnay, R. Gardner and B. Habecker for their valuable discussion and input into the project.

Author Contributions B.T.L. performed all *in vitro* experiments, time-lapse microscopy, ISP treatments, immunohistochemistry and data quantification. B.T.L., J.M.C. and Y.L.W. designed the peptides. M.A.D. and A.T. performed all surgical procedures. B.T.L., M.A.D., K.M.M. and A.T. performed behavioural testing. A.T., B.P.B. and K.X. helped perform the pull-downs. S.M.D. and S.K.-A. performed the CSPG signalling experiments. Y.S., S.K.-A., S.L. and S.A.B. contributed to experimental design and figure preparation. B.T.L. and J.S. designed all studies, analysed the data and wrote the paper. All authors discussed and helped prepare the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Readers are welcome to comment on the online version of the paper. The authors declare competing financial interests: details are available in the online version of the paper. Correspondence and requests for materials should be addressed to J.S. (JXS10@case.edu).

METHODS

In vitro DRG culture

DRG and satellite glia dissociation and culture. DRGs were harvested as previously described⁹. Briefly, DRGs were dissected from adult female Sprague–Dawley rats (Harlan) and incubated in a solution of collagenase II (200 U ml⁻¹, Worthington Biochemical Corporation) and dispase II (2.5 U ml⁻¹, Roche Diagnostics) in Ca²⁺/Mg²⁺ free Hank's balanced salt solution (HBSS-CMF, Invitrogen). Cells were centrifuged at 1,000–2,000 r.p.m., washed and gently triturated in HBSS-CMF three times. Dissociated DRGs were then resuspended in Neurobasal-A media supplemented with B-27, Glutamax, and penicillin/streptomycin (Invitrogen). DRGs were plated in Delta-T dishes (Fisher) at a density of 800 cells cm⁻² or on coverslips at a density of 1,000 cells cm⁻².

Dish preparation and time-lapse microscopy. Delta-T-cell culture dishes were prepared as previously described³¹. Culture dishes were rinsed with sterile water and then coated with poly-L-lysine (0.1 mg ml⁻¹, Sigma-Aldrich) overnight at room temperature, rinsed with sterile water, and allowed to dry. Aggrecan spot gradients were formed by allowing 2 µl of aggrecan solution (2 mg ml⁻¹ in HBSS-CMF, Sigma-Aldrich) to dry onto the culture surface. The surface of the dish was bathed in laminin solution (10 µg ml⁻¹ in HBSS-CMF, Invitrogen) for 3 h at 37 °C. The laminin bath was subsequently removed and cells were plated without allowing the surface of the dish to dry. For conditions with laminin alone, the aggrecan spots were excluded from the protocol.

Prior to time-lapse imaging, adult neurons were incubated at 37 °C for 4–6 days in Neurobasal-A media with either vehicle (water) or 2.5 µM peptide. Neurobasal-A media with HEPES (50 µM, Sigma-Aldrich) and either vehicle or 2.5 µM peptide was added to the cultures before imaging in a heated stage apparatus. Time-lapse images were acquired every 30 s for at least 1.5 h with a Zeiss Axiovert 405M microscope using a heated ×100 oil-immersion objective. Growth cones that extended normal to the spot rim were chosen for analysis. We tracked and charted the behaviour of growth cones in our *in vitro* assay with Metamorph software. *n* = 19 laminin, 24 control, 24 ISP, 26 ISP. For delayed treatment, *n* = 14 ISP and ChABC.

PTPσ concentration quantification. High-magnification images of PTPσ expression (R&D Systems, 1:100) in growth cones, either motile on uniform substrates of laminin (5 µg ml⁻¹) or dystrophic within the CSPG gradient were analysed with ImageJ. Both the growth cone and axon were manually traced and the mean pixel intensity was calculated. All images were taken using identical settings.

CSPG gradient crossing assay. CSPG gradients were prepared as described previously⁹. Glass coverslips (1.6 mm, Fisher Scientific) coated with poly-L-lysine and nitrocellulose were spotted with a 2 µl solution of aggrecan (0.7 mg ml⁻¹) and laminin (5 µg ml⁻¹) in HBSS-CMF (4 spots per coverslip). After the spots were allowed to dry, the coverslips were incubated with laminin (5 µg ml⁻¹) in HBSS-CMF at 37 °C for 3 h. Dissociated DRG neurons were plated at a density of 1,000 cells cm⁻² in Neurobasal-A supplemented with B27, Glutamax and penicillin/streptomycin and incubated for 5 days at 37 °C. For peptide experiments, appropriate concentrations of peptide were added to the media at the time of plating. For ChABC experiments, 0.1 U ml⁻¹ ChABC (Seikagaku) was added to coverslips for 2 h after the laminin bath before cell plating.

At 5 days, cultures were fixed in 4% paraformaldehyde in PBS for 30 min. After several rinses in PBS, the coverslips were incubated in blocking solution (5% normal goat serum or normal donkey serum, 0.1% BSA, and with or without 0.1% Triton X-100 in PBS) for 1 h at room temperature and then incubated overnight at 4 °C in primary antibody. Anti-βIII-tubulin (1:500; Sigma-Aldrich), anti-CS56 (1:500; Sigma-Aldrich), anti-PTPσ (1:100, R&D systems), anti-GFP (1:500, Invitrogen), anti-S100 (1:1,000, Sigma-Aldrich), anti-GFAP (1:5,000, DAKO), anti-NgR (1:100, Millipore) and anti-LAR (1:100, Santa Cruz) were used as primary antibodies. Coverslips were rinsed several times in PBS and then incubated in the appropriate secondary antibody (Molecular Probes) overnight at 4 °C. Coverslips were rinsed with PBS again, and mounted on glass slides in Citifluor (Ted Pella) mounting medium. Specimens were examined using a Leitz Orthoplan 2 fluorescence microscope.

Wedge identification. The intracellular domains of rat PTPσ (Protein Data Bank (PDB) accession 2FH7) and human LAR (PDB accession 1RPM) were visualized using UCSF Chimera 1.6.1. The wedge domains of human, rat and mouse PTPσ and PTPδ were identified by BLAST alignment with the known wedge domain of LAR¹⁴.

Peptide preparation. Peptides were synthesized commercially with C-terminal amidation (Genscript), and purity was assessed as >98% by mass spectrometry. Lyophilized peptides were dissolved in sterile water and stored at –80 °C until use. Peptide sequences are as follows: ISP, NH₂-GRKKRRQRRRCDMAEHMERLKA NDSLKLSQEYESI-NH₂; human ISP, NH₂-GRKKRRQRRRCDMAEHMERLKA NDSLKLSQEYESI-NH₂; ISP no TAT, NH₂-DMAEHMERLKA NDSLKLSQEYESI-NH₂; ILP, NH₂-GRKKRRQRRRCDLADNIERLKANDGLKFSQEYESI-NH₂; IDP, NH₂-GRKKRRQRRRCDELADHIERLKANDNLKFSQEYESI-NH₂; IMP, NH₂-DL LQHTQMCKAEGYGFKEEYESGRKKRRQRRRC-NH₂; scrambled ISP, NH₂-G

RKKRRQRRRCIREDDSLMLYALAEKKESNMHES-NH₂; TAT, NH₂-GRKKRRQRRRC-NH₂.

Quantification. All quantification was done blind. The number of βIII-tubulin-positive axons crossing the CSPG gradient (visualized using the CS56 antigen) was counted and divided by the total number of neuronal cell bodies contained within each gradient. Each gradient was counted as an individual data point. *n* = 112 control, 43 ChABC, 20 scrambled ISP, 57 ILP, 18 human ISP, 18 IDP, 16 IMP, 16 ISP no TAT, 16 TAT. For ISP, *n* (µm) = 16 (10), 16 (5), 34 (2.5), 49 (1.25), 27 (0.625, 0.25, 0.025, 0.0025).

For analysis of PTPσ concentration, PTPσ-stained neuronal growth cones and axons either entering the gradient or growing on uniform laminin were manually traced using ImageJ and the average pixel intensity was calculated.

Astrocyte preparation. Astrocytes were harvested from postnatal day 0/1 (P0/1) rat cortex as previously described³¹. Cortices were finely minced and treated with 0.5% trypsin in EDTA. Cells were plated in DMEM/F12 (Invitrogen) with 10% FBS (Sigma-Aldrich) and 2 mM Glutamax on T75 flasks coated with poly-L-lysine and shaken to remove non-adherent cells. Astrocytes were allowed to mature in culture for at least 4 weeks, and used within 2 weeks of maturity. Astrocytes were harvested with trypsin and plated at a density of 12,500 cells cm⁻².

Adhesion assay. Poly-L-lysine-coated glass coverslips were uniformly coated with a mixture of aggrecan (25 µg ml⁻¹) and laminin (1 µg ml⁻¹) for 3 h at 37 °C. DRG neurons were plated on coverslips (as described earlier) and incubated for 5 days at 37 °C. Cultures were then removed from the incubator and placed on a rotary shaker for 15 min at 80 r.p.m. Control coverslips were removed from the incubator, but not placed on the shaker. After shaking, the supernatants were immediately collected, centrifuged, re-suspended in 20 µl Neurobasal-A media, and placed on ice. The number of neurons released from each coverslip was counted with a haemocytometer. Coverslips were carefully fixed with 4% paraformaldehyde and stained for βIII-tubulin to visualize remaining neurons and axons (as described earlier). *n* = 28 control and ILP, 16 ISP.

DNA constructs and electroporation. mPTPσ in pECFP-N1 was a gift from A. M. Craig. Full-length mouse PTPσ with four fibronectin domains (BC052462) was subcloned into pEF1α-AcGFP1-N1 (Clontech) between NheI and HindIII. pEF1α-mPTPσ-AcGFP1-N1 was electroporated into adult DRG neurons with the Amaxa Rat Neuron Nucleofector Kit (Lonza) using the manufacturer's instructions.

PTPσ pulldown

Biotinylated peptide pulldown. For pulldown experiments, we used the Pierce Pull-Down Biotinylated Protein:Protein Interaction Kit (Thermo Scientific 21115). 100 µg ml⁻¹ of biotinylated-peptide (Genscript) was incubated overnight on an orbital shaker at 4 °C with streptavidin beads. After incubation, extra biotin was added and allowed to incubate overnight to ensure the binding of all streptavidin. After three washes with TBS, either recombinant GST-tagged PTPσ ICD (D1/D2 500 ng, Sigma, D1 500 ng, Abcam), spinal cord lysate from either wild-type or PTPσ-null mice, or brain lysate from an adult female Sprague–Dawley rat. Neural tissue was quickly extracted and flash frozen with liquid nitrogen. The tissue was homogenized in tissue homogenization buffer (20 mM Tris, 0.5 mM EDTA, 0.5 mM EGTA and 8% sucrose, pH 7.4) and 1:500 protease inhibitor cocktail (Abcam) on ice. The lysate was centrifuged at 13,000 r.p.m. for 20 min before addition to the beads. One-hundred and fifty microlitres of each lysate was added to the beads and allowed to incubate overnight at 4 °C. After three washes, beads were incubated with elution buffer for 10 min at room temperature. Beads were then centrifuged at 12,000 r.p.m. to collect eluted lysate.

SDS-PAGE and western blot. Thirty microlitres of the pulldown material with 4× Laemmli Buffer was boiled for 10 min at 100 °C and loaded into 7.5% TBX Mini-Protein Gels (Bio-Rad 456-1029) with Bio-Rad Precision Plus Protein Standard (Bio-Rad 161-0374). The gel was run for about 1.5 h at 100 V in 1× Tris Glycine SDS Buffer. Gels were stained with Sypro Ruby Red (Sigma-Aldrich) to visualize proteins within the gel. Transfer occurred overnight at 15 V using a PVDF membrane. We blocked at least 2 h in 5% milk powder, 0.1% Tween-20 before overnight incubation with an antibody against GST (Cell Signaling), PTPσ ICD (1:100, Abnova), PTPσ ECD (1:1,000, Abcam), anti-NgR (1:500, Millipore) or anti-LAR (1:1,000, R&D Systems). We washed five times for 5 min in 1× PBS-0.1% Tween-20 before blocking overnight with a horseradish peroxidase (HRP)-conjugated secondary (1:1,000). The blot was developed using a chemiluminescence substrate (Thermo Sci) after five times five 1× PBS/0.1% Tween-20 washes. All experiments were repeated >3 times.

CSPG signalling

Plating SH-SY5Y cells on laminin and CSPG substrates. The SH-SY5Y neuronal cell line (ATCC) was grown in Hyclone Dulbecco's modified Eagle's high-glucose media (GE Healthcare Life Sciences, SH-30081.02) supplemented with 4 mM L-glutamine, 1 mM sodium pyruvate, 1% penicillin/streptomycin/neomycin (PSN) and 10% heat-inactivated fetal bovine serum (Invitrogen). SH-SY5Y cells were plated at an initial density of 12,000 cells cm⁻² onto tissue culture surfaces for

4 days under different conditions, including (1) laminin, (2) laminin plus CSPG, (3) laminin plus CSPGs pre-treated with ChABC, (4) laminin plus ISP in the media, (5) laminin plus CSPG plus ISP in the media. Tissue culture dishes were coated with laminin ($2 \mu\text{g ml}^{-1}$, Sigma, L2020) and/or CSPG ($15 \mu\text{g ml}^{-1}$, Millipore, cc117) for 3 h at room temperature. Of note, CSPGs used in this study contained a mixture of neurocan, phosphacan, versican and aggrecan. Where appropriate, ChABC (0.1 U ml^{-1} , Sigma, C3667-10UN) was added with laminin plus CSPG mixture to tissue culture surfaces for 1 h and incubated at 37°C during coating and before cell plating. In the ISP condition, cells were pre-treated with ISP ($2.5 \mu\text{M}$) for 30 min.

Immunoblotting. Cells were harvested from culture plates 4 days after cell plating and homogenized in RIPA buffer (Thermo Fisher) containing SigmaFast Protease Inhibitor (Sigma). A total of 30–50 μg protein was loaded into the gel and then transferred to a nitrocellulose membrane (Bio-Rad). The membranes were then blocked in 5% non-fat milk in Tween Tris buffered saline (TBST) for 1 h and incubated overnight at 4°C with P-p44/42 MAPK (Cell Signaling, Rabbit 1:500) diluted in the blocking solution. The membranes were washed and incubated with HRP-conjugated goat anti-rabbit antibodies (1:4,000, Biorad). Membranes were then incubated in ECL plus immunoblotting detection reagents (Thermo Scientific Pierce) according to the manufacturer's specifications. Blots were then stripped of their primary and secondary antibodies for 30 min in 0.2 M NaOH and re-probed with primary antibody p44/42 MAPK (Erk1/2) (Cell Signaling, Rabbit 1:1,000) overnight followed by incubation with secondary antibody and ECL.

Quantification. Immunoreactive bands were quantified using AlphaEaseFC (FluorChem, 8900). The ratio of phosphorylated (p)Erk1/2 to total (t)Erk1/2 for each condition was calculated. Data are reported as means \pm s.e.m., and $P \leq 0.05$ was considered significant. Statistical analyses of intensity measurements were tested by one-way ANOVA comparing conditions followed by post-hoc pairwise multiple-comparison testing by the Holm–Sidak method.

Animals and contusive SCI

Animals. Adult female Sprague–Dawley rats (225–250 g) were obtained from Harlan. All procedures were approved by the Institutional Animal Care and Use Committees. PTP σ -null mice were provided by M. Tremblay. All PTP σ -null experiments were performed in collaboration with Y. Shen.

Contusive SCI. Briefly, adult female Sprague–Dawley rats (230–250 g) were obtained from Harlan and acclimated to the animal resource centre, behaviour analysis chambers and handlers. Rats were injected intraperitoneally with ketamine (60 mg kg^{-1}) and xylazine (10 mg kg^{-1}). The musculature was cut from T7–T9 and the dorsal surface of T8 was exposed by laminectomy. The vertebral column was stabilized by clamping the T7 and T9 vertebral bodies with forceps fixed to the base of an Infinite Horizon Impact Device. The animals were situated on the platform, and the 2.5 mm stainless steel impactor tip was positioned over the midpoint of T8 and impacted with 250 kdyn force. The overlying musculature was closed using suture, the skin was closed using wound clips and the animals were treated with Marcaine at the incision site. The force/displacement graph was used to monitor impact consistency and any animals that exhibited an abnormal impact graph or greater than 10% deviation from 250 kdyn were immediately excluded from the study.

After surgery, pain was monitored and animals were treated with intramuscular buprenorphine at signs of discomfort. In addition, manual bladder expression was performed 2–3 times daily for 2–3 weeks until a voiding reflex returned and animals could leak urine. Five vehicle animals and two ILP animals were removed from the study and euthanized due to bladder infections and other serious ailments. One ISP-treated animal was removed due to a bladder stone.

Dorsal column crush injury. Dorsal column crush was performed similarly to as described previously³¹. Rats were anaesthetized with inhaled isoflurane gas (2%) for all surgical procedures. A T1 laminectomy was performed to expose the dorsal aspect of the C8 spinal cord segment. Durotomies were made bilaterally 0.75 mm from midline with a 30-gauge needle. A dorsal column crush lesion was then made by inserting Dumont no. 3 jeweller's forceps into the dorsal spinal cord at C8 to a depth of 1.0 mm and squeezing the forceps, holding pressure for 10 s and repeating two additional times. The muscle layers were sutured with 4-0 nylon suture, and the skin was closed with surgical staples. Animals were perfused at 14 days after injury.

Systemic peptide treatment. All randomization and peptide treatments were prepared by a blinded laboratory member not associated with behavioural analyses. First, a vehicle solution of 1.25 ml DMSO in 23.75 ml sterile saline was prepared for each animal. Next, appropriate peptide was added to each of the vehicle solutions where applicable so that the final peptide concentration of each solution was $5 \mu\text{M}$. Each drug solution was then aliquotted into 50 individual 1.5 ml Eppendorf tubes, each corresponding to a single animal's daily dose, and frozen at -20°C . All peptides were randomized and blinded to both the animal treatment administrator and separate behavioural analyser. At 24 h after SCI and each morning thereafter for 49 consecutive days, animals were given a 500 μl subcutaneous injection of the

appropriate blinded treatment into the back above the lesion ($n = 21$ vehicle, $n = 26$ ISP, $n = 10$ ILP). This experiment was carried out with five different cohorts of animals, with freshly synthesized peptide validated *in vitro* using the CSPG gradient assay and prepared for each cohort of animals. ILP treatments were only performed in the first and last cohort because we did not observe behavioural improvement with ILP treatment.

For the dose response, additional injured animals ($N = 5$ per group, cohort 4) were injected with $1/3 \times$ ISP ($3.6 \mu\text{g}$ per day), $1/2 \times$ ISP ($5.5 \mu\text{g}$ per day), $2 \times$ ISP ($22 \mu\text{g}$ per day), $3 \times$ ISP ($33 \mu\text{g}$ per day) or $4 \times$ ISP ($44 \mu\text{g}$ per day).

FITC-ISP peptide tracking. FITC-ISP was synthesized by Genscript with FITC conjugated to the N-terminus TAT domain. A single 500 μl injection of either 10 μM FITC-ISP in 5% DMSO plus saline or vehicle was given to animals subcutaneously into the back. At 1 h after injection, spinal cords were immediately removed and snap frozen. Twenty-micrometre-thick coronal sections of unfixed spinal cord tissue were collected on a cryostat and immediately imaged on a Leitz Orthoplan 2 fluorescence microscope.

Methysergide. A subset of ISP- and vehicle-treated animals were injected with methysergide (5 mg kg^{-1} , Sigma-Aldrich) intraperitoneally once a day for 3 days at 14 weeks after SCI ($n = 8$ vehicle, $n = 12$ ISP). Gridwalk and metabolic cage analyses were performed on days 0 and day 3, while BBB scoring was conducted daily. In addition to the full population, the animals whose behaviour started above or below a threshold level of two standard deviations above vehicle mean (see Fig. 4f) were plotted separately as responding and non-responding.

Behavioural analysis

Open field BBB. All behaviour analyses were conducted by two blinded observers separate from the researcher performing the daily dosing. Each animal was tested on days 1, 4, 7 and weekly thereafter until week 11. Animals were allowed to freely roam on an open field while being observed by two Ohio State Spinal Cord Injury Course expertly trained observers and scored according to the BBB guidelines²⁷. Any animal with a BBB score of greater than 1 at day 1 was removed from the study. Data were quantified as the average of the two hind limbs, compiled, and graphed.

Gridwalk. The gridwalk test was performed at 12 weeks after SCI. Animals were allowed to freely roam on a $75 \text{ cm} \times 40 \text{ cm}$ raised grid (2.5 mm thick wires, 2 cm gaps between wires) for 5 min while their progress was tracked with an overhead camera and quantified as total distance travelled (Ethovision). Foot faults were counted manually by a blinded observer and quantified as total number of hindlimb faults per metre.

Thermal hyperalgesia. Hyperalgesia analysis was performed at 12 weeks after injury by a blinded observer as published previously³². Animals were given 30 min to acclimate to the plexiglass cage before testing (Ugo Basile). The infrared radiation source (intensity = 58) was carefully placed under each hindpaw. Time to withdrawal was recorded as an average of five trials on each paw, with the longest and shortest time removed.

Mechanical allodynia. The Von Frey hair protocol for the hind paw was adapted from the Ohio State University Spinal Cord Injury Program. Briefly, animals satisfying the weight-bearing criteria were acclimated to the Von Frey testing boxes for at least 15 min. A total of ten trials were performed starting with the 5.18 monofilament while animals were distracted with a treat. The monofilament was tested on the plantar surface of the centre of the paw between the foot pads. A positive response was recorded if an animal withdrew its paw when the monofilament was presented. At least 30 s elapsed between each trial for the same hind paw. Positive responses led us to test with progressively smaller monofilaments. Conversely, negative responses led us to test with progressively larger monofilaments until monofilament 6.10. The threshold value was defined as the lowest monofilament level at which 50% or more of the trials resulted in a positive withdrawal.

Metabolic cage micturition analysis. At 6 weeks and 12 weeks after SCI, animals were placed in a metabolic cage (Baintree Scientific) for measurement of voiding patterns. The voided urine was measured continuously via a force transducer/strain gauge (Grass Technologies) and plotted in Spike2 (Cambridge Electrical Design, sampled at 20 Hz). Animals were kept in this cage for 16 h with ample water and food during the period of urine collection and measurement. The criteria for the micturition pattern analysis included void frequency (voids per h) and the void volume (ml per void). The total volume of expelled urine was not included because of variations in water intake between individual animals. $n = 11$ naive animals.

Urodynamics. Terminal urodynamic recordings were performed similarly to as described previously³³. Briefly, rats were anaesthetized at 14 weeks after SCI with 0.8 g kg^{-1} urethane delivered subcutaneously. A polyethylene-50 catheter was carefully inserted through the urethra into the bladder for delivery of saline. Fine wire electrodes ($76.2 \mu\text{m}$) (0.003" diameter Teflon-insulated silver wire, A-M Systems) were inserted percutaneously via the vagina on both sides of the urethra to monitor the EUS electromyography (EMG) activity. The electrodes were connected to a preamplifier (HZP; Grass-Technologies), which was connected to an amplifier (Grass-Technologies) with low- and high-pass frequency filters at 30 Hz and 3 kHz,

respectively, and signal was sampled at a rate of 10 kHz (Power 1401, Spike2; Cambridge Electronic Design). Continuous cystometrograms (CMGs) were collected using constant infusion (6 ml h^{-1}) of room temperature saline (Aladdin-1000 single syringe infusion pump; World Precision Instruments) through the catheter into the bladder to elicit repetitive voids. The bladder pressure was recorded via the same catheter used for saline infusion, using a pressure transducer (Grass Technologies) connected to the recording system and sampled at a frequency of 2 kHz. Animals that received methysergide did not receive urodynamic analysis. $n = 11$ vehicle, 15 ISP, 6 ILP.

Immunocytochemistry and tracing

Perfusion and sectioning. To obtain spinal cord sections, rats were transcardially perfused with ice-cold 4% paraformaldehyde in PBS, and the spinal cords were dissected out. After the tissue was postfixed in 4% paraformaldehyde overnight at 4°C and cryoprotected with 30% sucrose, spinal cords were frozen in OCT mounting media and sectioned on a Hacker cryostat at a thickness of 20 μm .

Immunocytochemistry. Mounted sections were washed three times with PBS followed by blocking in 5% normal goat serum (NGS) and 0.1% bovine serum albumin (BSA) in PBS. 0.1% Triton X-100 was added to the blocking buffer depending on the antigen used. After blocking, sections were incubated in primary antibody diluted in blocking buffer overnight at 4°C . Primary antibodies used were mouse anti-NeuN (1:100, Chemicon), anti-GFAP (1:1,000, Dako), mouse anti-ED1 (Chemicon, 1:100) rabbit anti-5-HT (1:500, Immunostar), and anti-neurofilament (1:500, Sigma-Aldrich). For *in vivo* PTP σ staining, sagittal 20 μm sections encompassing regions both rostral and caudal to the lesion were probed with anti-PTP σ antibody (1:500, Abnova). The next day, the sections were washed extensively with PBS and incubated in the appropriate secondary antibody or avidin substrate conjugated to AlexaFluor 488, 594 or 633 (1:500, Molecular Probes) overnight. After extensive washing, the sections were stained with DAPI (1:1,500 in PBS, Sigma-Aldrich), washed, coverslipped and viewed with a confocal microscope (Zeiss, Germany). Pixel intensity was measured on images taken on a standard fluorescent microscope (Leica) with a uniform exposure setting and analysed using ImageJ.

White matter analysis. Spared white matter analysis was conducted by EC staining as previously published³⁴. Briefly, room temperature 20 μm sections were placed in fresh acetone for 10 min, removed and allowed to dry for 30 min. Sections were stained with freshly filtered EC solution (Sigma-Aldrich) for 30 min and washed in running tap water for 5 min. The stain was differentiated in 5% ferric ammonium sulphate (Sigma-Aldrich) for 15 min and again washed with running tap water for 5 min. The differentiation was completed with borax-ferricyanide solution (Sigma-Aldrich) for 10 min, briefly washed with running tap water and allowed to dry. Slides were dehydrated in 70%, 95% and 100% ethanol for 2 min each followed by xylene for 2 min. Slides were coverslipped with a VectaMount Permanent Mounting Medium (Vector Laboratories). $n = 10$ vehicle, 14 ISP.

Lesion volume. Spinal cord sections stained with GFAP were analysed for lesion volume. After staining, sections were digitalized with a Leica SCN 400 Slide Scanner. Fifteen 20 μm coronal sections (one of every ten serial sections was stained, 200 μm between sections) both rostral and caudal of lesion epicentre were analysed for a total of 6 mm of spinal tissue. Sections were traced in ImageJ for volume calculation. $n = 9$ vehicle, 14 ISP.

5HT analyses. Coronal sections of lumbar spinal cord were analysed for 5HT intensity. The staining and imaging settings were uniform for all images. High exposure settings were used to maximize the signal to noise ratio, allowing for large patterns of 5HT expression to be visualized. ImageJ threshold analysis was used to eliminate all background from each section, leaving only the patterns of 5HT expression. The

area of innervation and per cent coverage of the grey matter was identified by analysing all particles in the grey matter at threshold intensity. Eight 20 μm sections, (corresponding to 1.6 mm of lumbar tissue) were analysed in ImageJ and averaged, with the highest and lowest intensity removed. $n = 13$ vehicle, 18 ISP.

Tracing. Cortical spinal tract labelling was performed as published previously³⁵. Ten per cent BDA (Molecular Probes) was injected into 16 locations in the rat motor cortex at 12 weeks after injury. Animals were killed 2 weeks after labelling. BDA labelling was visualized using an avidin-biotin peroxidase incubation followed by diaminobenzidine and H_2O_2 (Vector Labs).

Statistical analysis. All statistical analyses were performed using Graphpad Prism. All results are presented as mean \pm s.e.m. Sample sizes were initially determined using statistical software to calculate the minimum total required number of animals or assays. All reported groups are above the minimum calculated sample size.

In vitro. For PTP σ intensity measurements, gradient crossings and adhesion assays, all statistical analysis was performed with one-way ANOVA between all groups in each individual experiment.

In vivo. For void frequency, void volume and gridwalk data sets, D'Agostino-Pearson and Shapiro-Wilk tests were first performed to determine if any individual data set abided to a normal, Gaussian distribution. Naive, vehicle control and ILP all passed the normality test, while ISP failed in all. Next, we performed ROUT analysis to identify outliers. No values were excluded from analysis, although a single ILP void frequency data point was identified as an outlier. Statistical analysis on these behaviours was performed with a one-way ANOVA and post-hoc Kruskal-Wallis test.

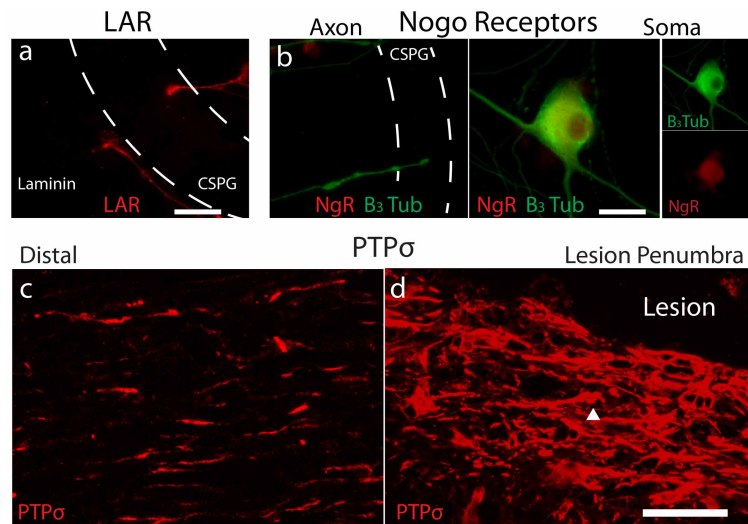
For BBB, we performed a repeated measures two-way ANOVA to compare vehicle, ILP and ISP. No normality tests were performed because BBB is a nonlinear scale. Post-hoc analysis at individual time point was performed with Tukey's test.

After methysergide treatment, BBB analysis was performed with repeated measures two-way ANOVA. For both gridwalk and void frequency, we compared pre-methysergide to post-methysergide treatment with a two-way ANOVA. The analysis between responding and non-responding animals was performed separately.

Anatomy. 5HT area and per cent grey matter coverage, the average value over the eight lumbar sections of ISP and vehicle control were compared with a Student's *t*-test.

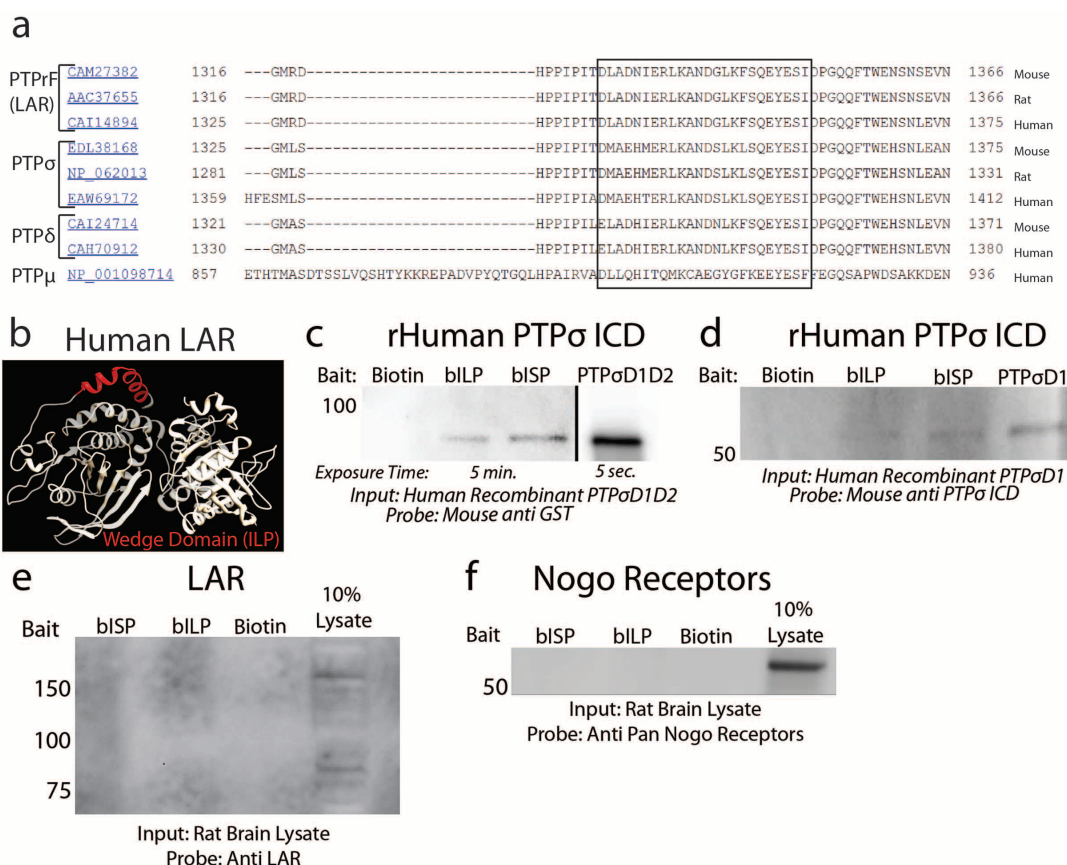
Regression analysis. Regression analysis was performed by creation of a matrix of behavioural and anatomical scores for each animal. A separate matrix was created for vehicle and ISP treatment. Regression values (*r*) were identified by comparing two individual variables. The Pearson or Spearman coefficients were identified for parametric and non-parametric data sets, respectively.

31. Horn, K. P., Busch, S. A., Hawthorne, A. L., van Rooijen, N. & Silver, J. Another barrier to regeneration in the CNS: activated macrophages induce extensive retraction of dystrophic axons through direct physical interactions. *J. Neurosci.* **28**, 9330–9341 (2008).
32. Hargreaves, K., Dubner, R., Brown, F., Flores, C. & Joris, J. A new and sensitive method for measuring thermal nociception in cutaneous hyperalgesia. *Pain* **32**, 77–88 (1988).
33. Cheng, C. L. & de Groat, W. C. The role of capsaicin-sensitive afferent fibers in the lower urinary tract dysfunction induced by chronic spinal cord injury in rats. *Exp. Neurol.* **187**, 445–454 (2004).
34. Jakeman, L. B. in *Animal Models of Acute Neurological Injuries II* (eds Chen, J. Xu, X.-M., Xu, Z. C. & Zhang, J. H.) 417–442 (Springer, 2012).
35. Weidner, N., Ner, A., Salimi, N. & Tuszynski, M. H. Spontaneous corticospinal axonal plasticity and functional recovery after adult central nervous system injury. *Proc. Natl Acad. Sci. USA* **98**, 3513–3518 (2001).



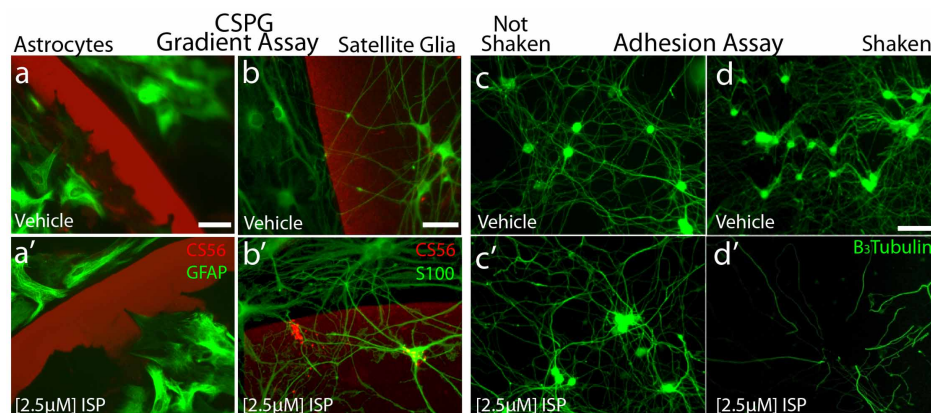
Extended Data Figure 1 | CSPG receptors LAR and NgR *in vitro* and PTPσ *in vivo*. **a**, LAR expression in motile (left, on laminin) and immobilized (right, within CSPG gradient) growth cones. **b**, Nogo receptors in a soma and

axons. Scale bars, 20 μm. **c, d**, PTPσ expression in the spinal cord 14 days after dorsal column crush injury. The arrowhead pointing upwards represents a labelled structure with dystrophic morphology. Scale bar, 50 μm.



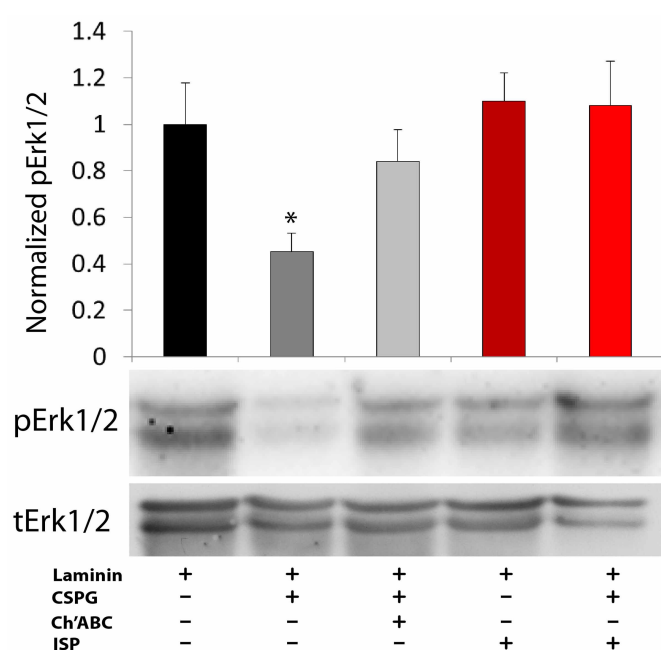
Extended Data Figure 2 | LAR structure, sequence alignment and pulldown analysis. **a**, BLAST alignment of the known sequences of mouse, rat and human LAR, PTPσ, PTPδ and PTPμ. The wedge domain of each protein is aligned within the box. **b**, The tandem intracellular phosphatase domains of

human LAR with the previously characterized wedge domain (red)¹⁴. **c**, **d**, Pulldown of recombinant PTPσ with biotinylated (b)ISP or ILP. **e**, **f**, Eluted lysate after pulldown was probed with antibodies against either LAR or pan-NgRs. Input is 10% lysate control.

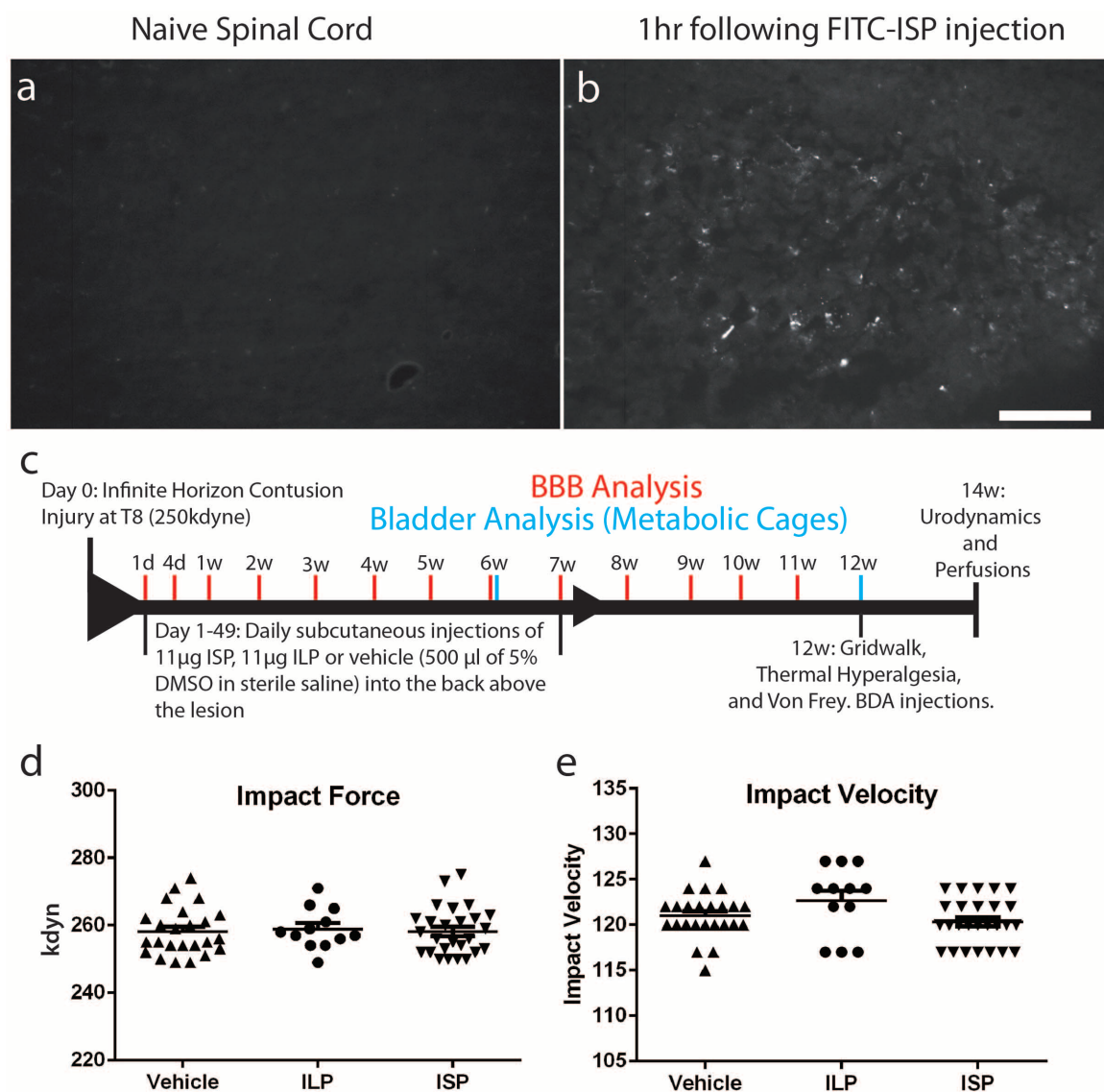


Extended Data Figure 3 | Astrocyte and satellite cell response to ISP and the adhesion assay. **a**, Purified GFAP-positive mature astrocytes (green) did not respond to ISP. **b**, S100-positive satellite glia were able to cross the gradient

of CSPG after ISP treatment. Scale bar, 50 μm . **c**, **d**, Response of neurons and axons upon a CSPG-rich substrate to agitation after ISP treatment. Scale bar, 50 μm .

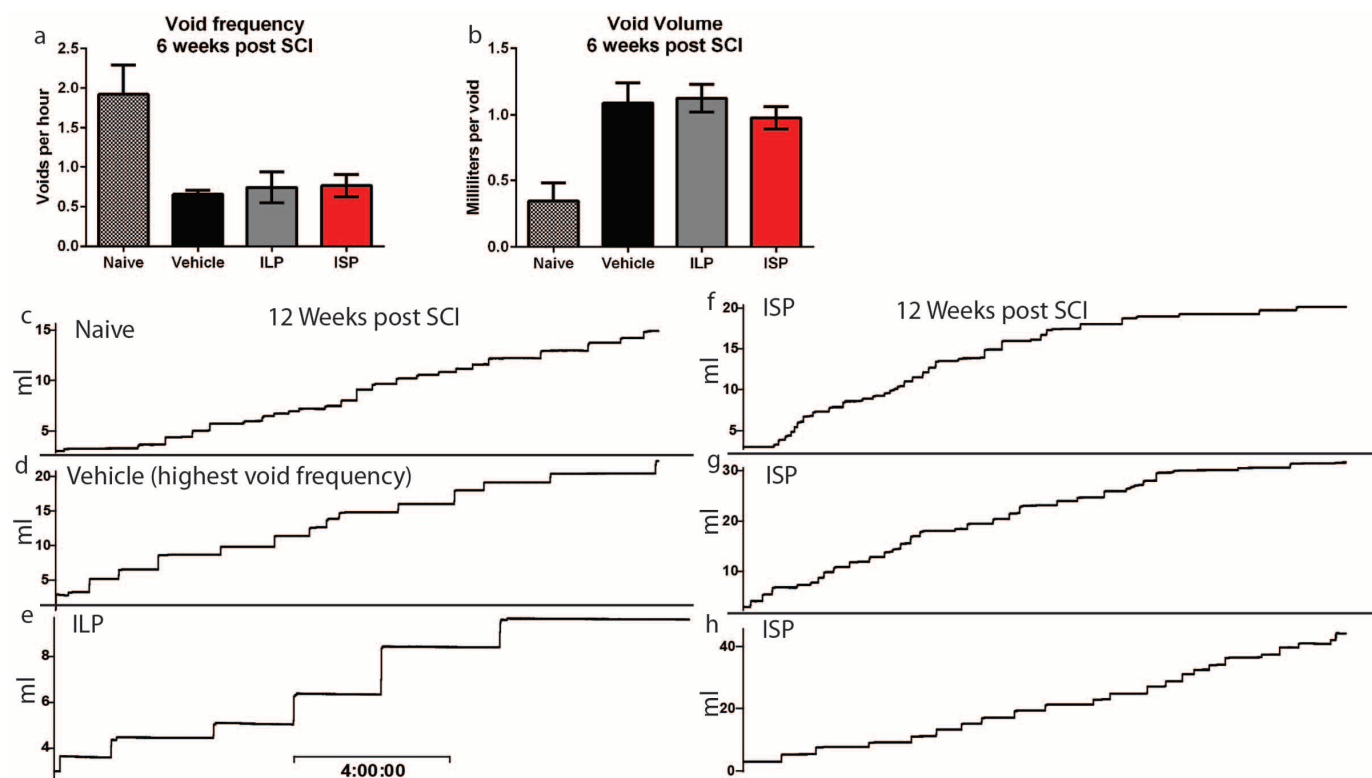


Extended Data Figure 4 | CSPG and ISP regulation of Erk1/2. Western blot analysis revealed a significant decrease in the phosphorylation ratio of Erk1/2 in SH-SY5Y cells plated on laminin (2 $\mu\text{g ml}^{-1}$) plus CSPGs (15 $\mu\text{g ml}^{-1}$) compared with laminin-only substrates, which was reversed by either pre-treatment with ChABC (0.1 U ml^{-1}) or 4 days of ISP treatment (2.5 μM). pErk1/2, phosphorylated Erk1/2; tErk1/2, total Erk1/2. Data normalized to laminin control. $N = 4$ independent experiments. $*P < 0.05$ versus all other conditions, one-way ANOVA, Tukey's post-hoc test.



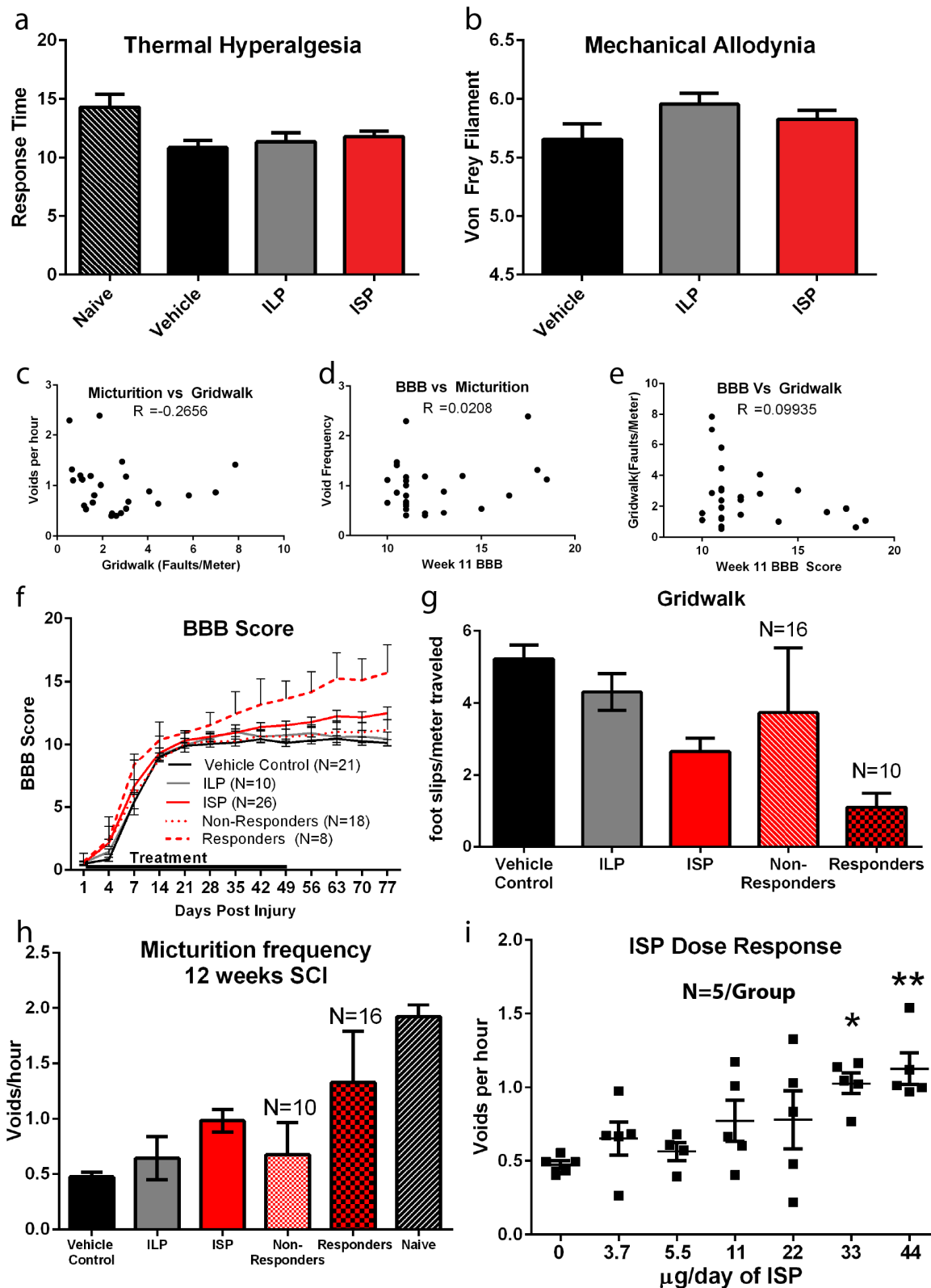
Extended Data Figure 5 | FITC-ISP *in vivo* and infinite horizon impactation.
a, b, Spinal cord 1 h after subcutaneous injections of FITC-ISP or vehicle. Scale bar, 100 µM. **c,** Experimental design and timeline for *in vivo* experiment.

d, e, The force and impactation velocity of all animals that received an infinite horizon contusive injury. All impactations are within 10% from the target force of 250 kdyn, with an average force of 258.2 for both ISP and vehicle.



Extended Data Figure 6 | Metabolic cage analysis at 6 weeks after injury.
a, b, Void frequency and average void volume at 6 weeks after injury. $n = 11$ naive, 21 vehicle, 10 ILP and 26 ISP. **c–h,** Representative smoothed metabolic

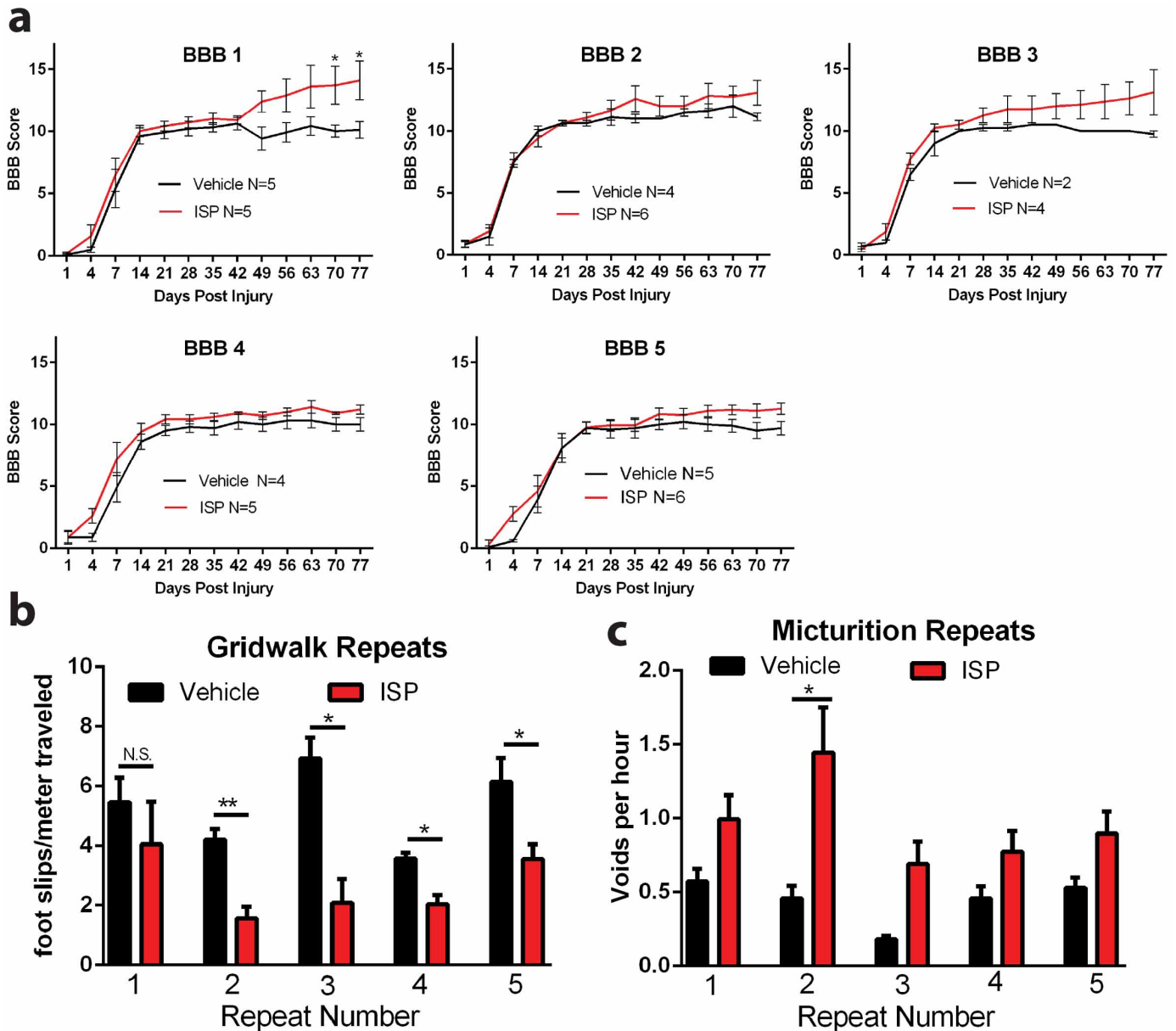
cage traces of a normal animal (**c**) and five treated animals (vehicle (**d**), ILP (**e**) and ISP (**f, h**) 12 weeks after injury. Void volume is plotted (in ml) as a function of time. Scale bar, 4 h.



Extended Data Figure 7 | Additional behavioural data and analyses.

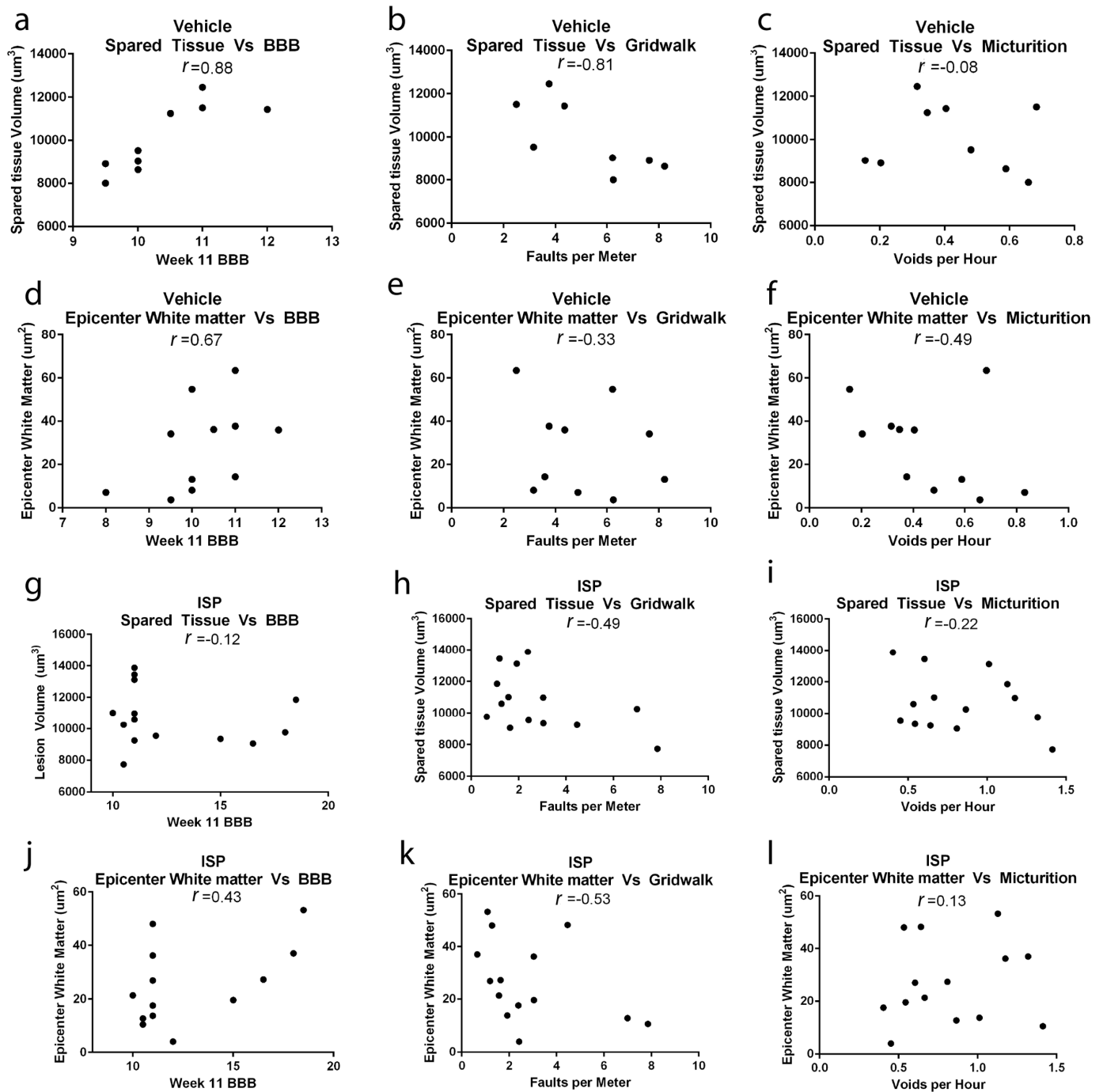
a, b, Average response to thermal (Hargrave's test) and mechanical (Von Frey test) stimuli at 12 weeks after injury ($n = 11$ naive, 21 vehicle, 10 ILP and 26 ISP for thermal, $n = 10$ vehicle, 4 ILP and 10 ISP for mechanical). **c–e**, Correlations between recovery of each individual motor behaviour in the ISP treatment group. **f–h**, For each behaviour, the ISP-treated animals that

recovered to two standard deviations relative to vehicle mean were separated and plotted as 'responders' while those that did not were plotted as 'non-responders'. The n for the responding and non-responding group for each behaviour is listed on the graph. **i**, An *in vivo* ISP dose-response plot in a single cohort of animals. A dose-dependent increase occurred in void frequency at 12 weeks after SCI.



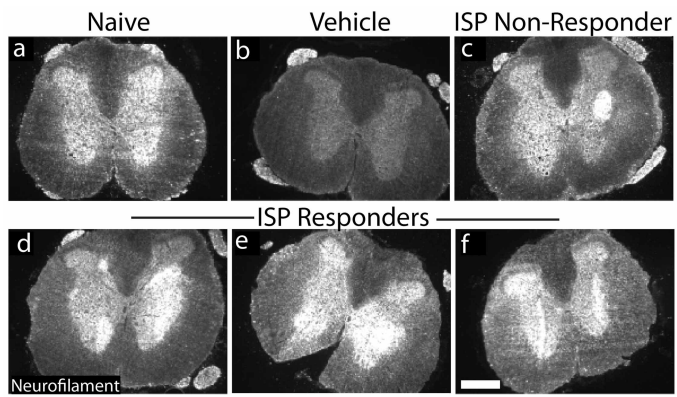
Extended Data Figure 8 | Five repetitions of *in vivo* experiments. a–c, The individual results of five repeats of *in vivo* experiments are plotted as individual cohorts of animals for BBB, gridwalk and void frequency. * $P < 0.05$, ** $P < 0.01$, repeated measures two-way ANOVA, Tukey's post-hoc test (BBB);

one-way ANOVA, Kruskal–Wallis post-hoc test (gridwalk and void frequency). Black indicates ISP versus control; red indicates ISP versus ILP. *n* for each condition is listed.



Extended Data Figure 9 | Correlation between spared tissue and behavioural recovery. a–l, Spared tissue volume (as measured by GFAP-positive tissue) or area of spared white matter at the epicentre (as measured by eriochrome cyanine staining) were plotted against behavioural scores for

vehicle- (a–f) and ISP-treated (g–l) animals. Pearson's correlation coefficient (r value) is reported for each comparison. Only animals whose spinal cord was processed and cut coronally were included in the analysis.



Extended Data Figure 10 | Neurofilament staining at lumbar levels.
a–f, Representative caudal neurofilament expression in lumbar spinal cord. Responders are ISP animals that demonstrated functional improvement (see Fig. 3n). All images were taken using identical settings. Scale bar, 500 μ m.

Towards a therapy for Angelman syndrome by targeting a long non-coding RNA

Linyan Meng^{1*}, Amanda J. Ward^{2*}, Seung Chun², C. Frank Bennett², Arthur L. Beaudet¹ & Frank Rigo²

Angelman syndrome is a single-gene disorder characterized by intellectual disability, developmental delay, behavioural uniqueness, speech impairment, seizures and ataxia^{1,2}. It is caused by maternal deficiency of the imprinted gene *UBE3A*, encoding an E3 ubiquitin ligase^{3–5}. All patients carry at least one copy of paternal *UBE3A*, which is intact but silenced by a nuclear-localized long non-coding RNA, *UBE3A* antisense transcript (*UBE3A-ATS*)^{6–8}. Murine *Ube3a-ATS* reduction by either transcription termination or topoisomerase I inhibition has been shown to increase paternal *Ube3a* expression^{9,10}. Despite a clear understanding of the disease-causing event in Angelman syndrome and the potential to harness the intact paternal allele to correct the disease, no gene-specific treatment exists for patients. Here we developed a potential therapeutic intervention for Angelman syndrome by reducing *Ube3a-ATS* with antisense oligonucleotides (ASOs). ASO treatment achieved specific reduction of *Ube3a-ATS* and sustained unsilencing of paternal *Ube3a* in neurons *in vitro* and *in vivo*. Partial restoration of UBE3A protein in an Angelman syndrome mouse model ameliorated some cognitive deficits associated with the disease. Although additional studies of phenotypic correction are needed, we have developed a sequence-specific and clinically feasible method to activate expression of the paternal *Ube3a* allele.

Phosphorothioate-modified chimeric 2'-O-methoxyethyl (2'-MOE) DNA ASOs ($n = 240$) were designed complementary to a 113 kilobase pair (kb) region of mouse *Ube3a-ATS* downstream of the *Snord115* cluster of small nucleolar RNAs (snoRNAs) (Fig. 1a). After nuclear hybridization of the ASO to the target RNA, RNase H cleaves the RNA strand of the ASO–RNA heteroduplex, resulting in subsequent RNA degradation by exonucleases¹¹. A high-throughput imaging screen identified ASOs that unsilenced the *Ube3a* paternal allele. Primary neurons from *Ube3a*^{+/YFP} (Pat^{YFP}) knock-in mice¹² were cultured and treated with ASO (15 μ M, 72 h), and we determined the fold increase of paternal yellow fluorescent protein (YFP)-tagged UBE3A (UBE3A–YFP) signal in NeuN (also known as Rbfox3)-positive cells (Fig. 1b). The non-targeting control ASO (Ctl ASO) had no effect on fluorescence (0.96 ± 0.01) whereas the positive control topoisomerase I inhibitor (topotecan, 300 nM) increased fluorescence (3.61 ± 0.00). ASO A and ASO B resulted in an increase in paternal UBE3A–YFP fluorescence of 2.11 ± 0.02 and 2.47 ± 0.03 , respectively (Fig. 1c). ASOs modulated RNA expression in a dose-dependent manner with greater than 90% reduction of *Ube3a*^{YFP}-*ATS* (Fig. 1d, top) within 48 h of treatment (Fig. 1d, bottom).

Snrpn, *Snord116* and *Snord115* are processed from the same precursor transcript as *Ube3a-ATS* (Fig. 1a) and are critical genes in Prader–Willi Syndrome (PWS)¹³. Their expression was not affected by increasing the dose or time of ASO treatment (Fig. 1d, e). The ability to downregulate *Ube3a-ATS* without affecting *Snord116* expression can be attributed to a fast rate of *Snord116* splicing (approximately 30 min) relative to the length of time required for transcription of the 332 kb region between *Snord116* and the ASO-binding site (approximately 80 min) (Extended Data Fig. 1). While *Ube3a-ATS* ASOs did not affect expression of mature *Snord116* or its precursor, ASOs designed directly to *Snord116*

strongly reduced *Snord116* and the entire *Ube3a-ATS* precursor transcript (Extended Data Fig. 1).

ASO treatment (10 μ M, 24 h) specifically reduced *Ube3a-ATS* (1,000 kb) without affecting expression of five other long genes (*Nrxn3*, 1,612 kb; *Astn2*, 1,024 kb; *Pcdh15*, 828 kb; *Csmd1*, 1,643 kb; *Il1rapl1*, 1,368 kb), whereas topotecan (300 nM, 24 h), which acts by impairing transcription elongation¹⁴, strongly inhibited their expression (Fig. 1f).

Primary neurons from Pat^{YFP} mice treated with ASO (10 μ M, 72 h) or topotecan (300 nM, 72 h) resulted in biallelic UBE3A protein expression due to unsilencing of the paternal allele (Fig. 1g). Additionally, ASO treatment of primary neurons from *Ube3a*^{KO/+} (Angelman syndrome) mice¹⁵ achieved 66–90% of wild-type levels of UBE3A protein (Fig. 1h). ASO treatment (10 μ M) did not affect DNA methylation at the PWS imprinting centre (Fig. 1i). A sequence-matched ASO that was rendered unresponsive to RNase H by complete modification with 2'-MOE nucleotides (ASO, inactive) did not affect paternal UBE3A expression, indicating that reduction of the antisense transcript is required for paternal *Ube3a* unsilencing (Fig. 1g).

Although reduction of the antisense transcript was required, additional studies indicated that it was not sufficient for paternal *Ube3a* unsilencing. ASOs complementary to the region of *Ube3a*^{YFP}-*ATS* upstream of *Ube3a* (non-overlapping ASOs, $n = 15$) upregulated *Ube3a*^{YFP} RNA 7.4 ± 0.6 fold relative to untreated control neurons (Extended Data Fig. 2). ASOs complementary to the region of *Ube3a*^{YFP}-*ATS* located within the *Ube3a* gene body (overlapping ASOs, $n = 12$) only upregulated *Ube3a*^{YFP} RNA 1.7 ± 0.2 fold. Because both non-overlapping and overlapping ASOs reduced *Ube3a*^{YFP}-*ATS* to a similar level, a mechanism independent of the presence of the long non-coding RNA may have a role in *Ube3a* silencing.

Next, we tested whether central nervous system (CNS) administration of *Ube3a-ATS* ASOs unsilenced paternal *Ube3a* *in vivo*. A single intracerebroventricular (ICV) injection of ASO was administered into the lateral ventricle of adult Pat^{YFP} mice. The ASO treatment was generally well tolerated, despite transient sedation after surgery. No significant changes in body weight, expression of AIF1 (marker for microgliosis), or expression of GFAP (marker for astrocytosis) were observed 1 month after treatment (Extended Data Fig. 3). Four weeks after treatment, ASO A and ASO B reduced *Ube3a-ATS* RNA by 60–70% and upregulated paternal *Ube3a*^{YFP} RNA two- to five-fold in the brain and spinal cord (Fig. 2a). However, compared with *Ube3a*^{YFP/+} (Mat^{YFP}) mice, ASO treatment did not fully unsilence the paternal allele. *Ube3a*^{YFP} RNA in ASO-treated Pat^{YFP} mice was 30–40% of the level in Mat^{YFP} mice (Fig. 2a). Western blot quantification showed that UBE3A–YFP protein was upregulated in the cortex ($82 \pm 7\%$), hippocampus ($33 \pm 3\%$) and thoracic spinal cord ($73 \pm 33\%$) in ASO-A-treated Pat^{YFP} mice compared with Mat^{YFP} mice (Fig. 2b). No significant downregulation of *Snrpn*, *Snord116*, *Snord115*, or the sentinel long genes was observed, including any *Snord116* reduction in the hypothalamus (Fig. 2a, c and Extended Data Fig. 4).

¹Department of Molecular and Human Genetics, Baylor College of Medicine, and Texas Children's Hospital, Houston, Texas 77030, USA. ²Department of Core Antisense Research, Isis Pharmaceuticals, Carlsbad, California 92010, USA.

*These authors contributed equally to this work.

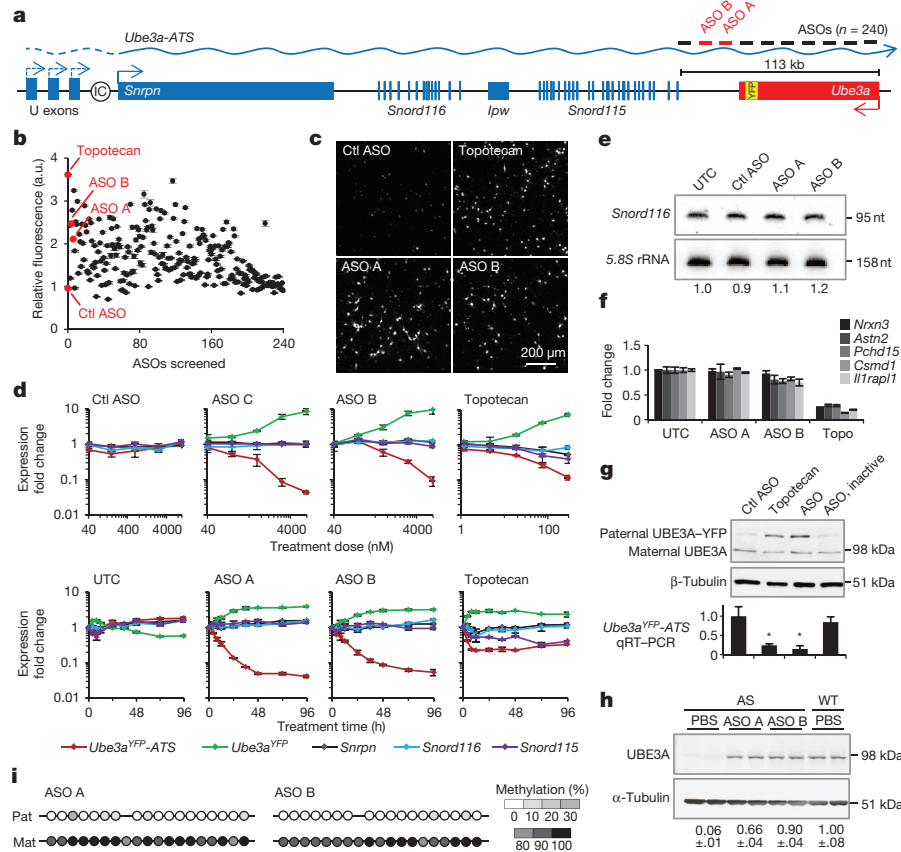


Figure 1 | Unsilencing of the *Ube3a* paternal allele by *Ube3a*-ATS-targeted ASOs in cultured mouse neurons. **a**, Schematic mouse *Ube3a* genomic locus. IC, imprinting centre. **b**, UBE3A-YFP fluorescence (arbitrary units (a.u.)) in ASO-treated primary neurons relative to untreated control (UTC). **c**, YFP fluorescent imaging of treated Pat^{YFP} neurons. **d**, Normalized mRNA levels in Pat^{YFP} neurons treated with increasing doses (top) or for increasing times (bottom). **e**, Northern blot of *Snord116* expression. *Snord116* intensity relative to 5.8S ribosomal RNA is quantified. nt, nucleotides. **f**, Normalized mRNA levels of long genes. Topo, topotecan. **g**, Western blot (top) and reverse transcription with quantitative polymerase chain reaction (qRT-PCR) (bottom) from Pat^{YFP} neurons. 'ASO, inactive' is a sequence-matched RNase H inactive ASO. * $P < 0.05$, two-tailed t -test, $n = 2$ per group, mean \pm absolute deviation. **h**, Western blot from wild-type (WT) or Angelman syndrome (AS) primary neurons. UBE3A signal intensity was quantified relative to α -tubulin. **i**, DNA methylation analysis of the PWS imprinting centre. The paternal allele was distinguished by the conversion of a CpG dinucleotide (CG > AA) in CAST.Ch7 mice.

After a single ASO dose, *Ube3a*-ATS reduction was sustained for 16 weeks in the CNS, and returned to basal expression by 20 weeks after treatment (Fig. 2d). Both the RNA and protein levels of paternal UBE3A-YFP were significantly higher than in PBS-treated mice at 2 to 16 weeks after treatment, and returned to the silenced state 20 weeks after treatment (Fig. 2d, e). No significant changes in *Snrpn*, *Snord115* or *Snord116* expression were observed (Fig. 2d). Immunostaining on brain sections 16 weeks after treatment further confirmed the long stability of the ASO and duration of paternal UBE3A protein expression (Extended Data

Fig. 5). This result is consistent with the long stability of other centrally administered ASOs that are chemically modified to resist intracellular nuclease degradation^{16,17}.

After ICV delivery, the ASO displayed widespread bilateral distribution throughout the brain, as demonstrated by immunostaining (Fig. 3a), and *in situ* hybridization confirmed the *in vivo* downregulation of *Ube3a*-ATS (Fig. 3b). UBE3A-YFP protein was expressed in ASO-positive cells (Fig. 3c). Increased UBE3A-YFP signal was detected in NeuN-positive cells throughout the brain (Fig. 3d and Extended Data Figs 6 and 7).

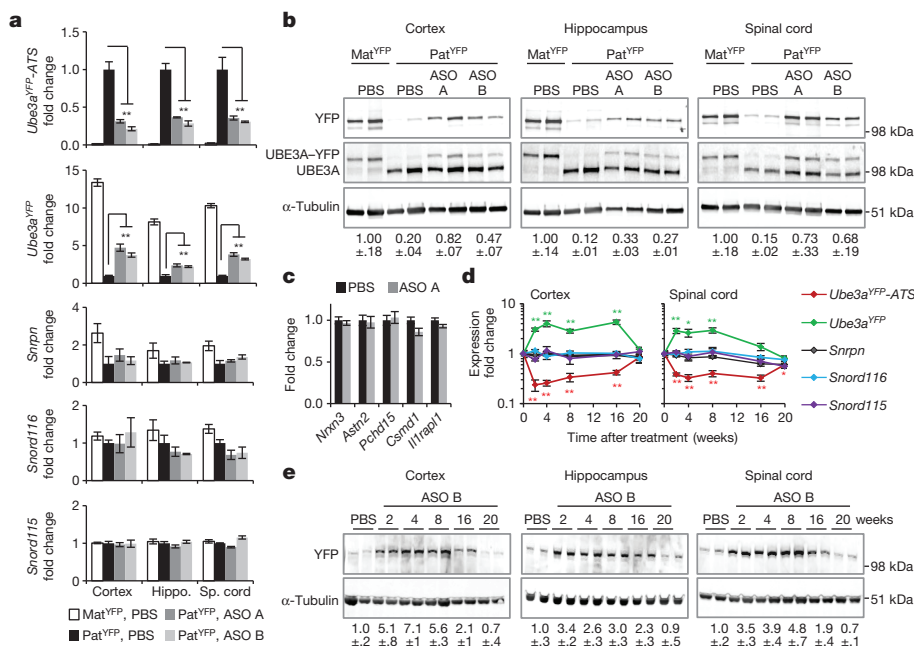


Figure 2 | A single administration of *Ube3a*-ATS ASOs resulted in paternal UBE3A unsilencing for 4 months. **a**, **b**, mRNA levels (**a**) and UBE3A-YFP protein (**b**) in cortex, hippocampus (Hippo.) and spinal cord (Sp. cord) 4 weeks after ICV injection of PBS or ASO in Pat^{YFP} mice. Mat mice are included for comparison. **c**, Normalized mRNA levels of long genes in the cortex. **d**, **e**, RNA levels (**d**) and UBE3A-YFP protein (**e**) in ASO-treated Pat^{YFP} mice 2–20 weeks after treatment. * $P < 0.05$, ** $P < 0.005$, two-tailed t -test, $n = 3$ –4 per group, mean \pm standard error of the mean (s.e.m.). For western blot quantification, YFP signal intensity was calculated relative to α -tubulin.

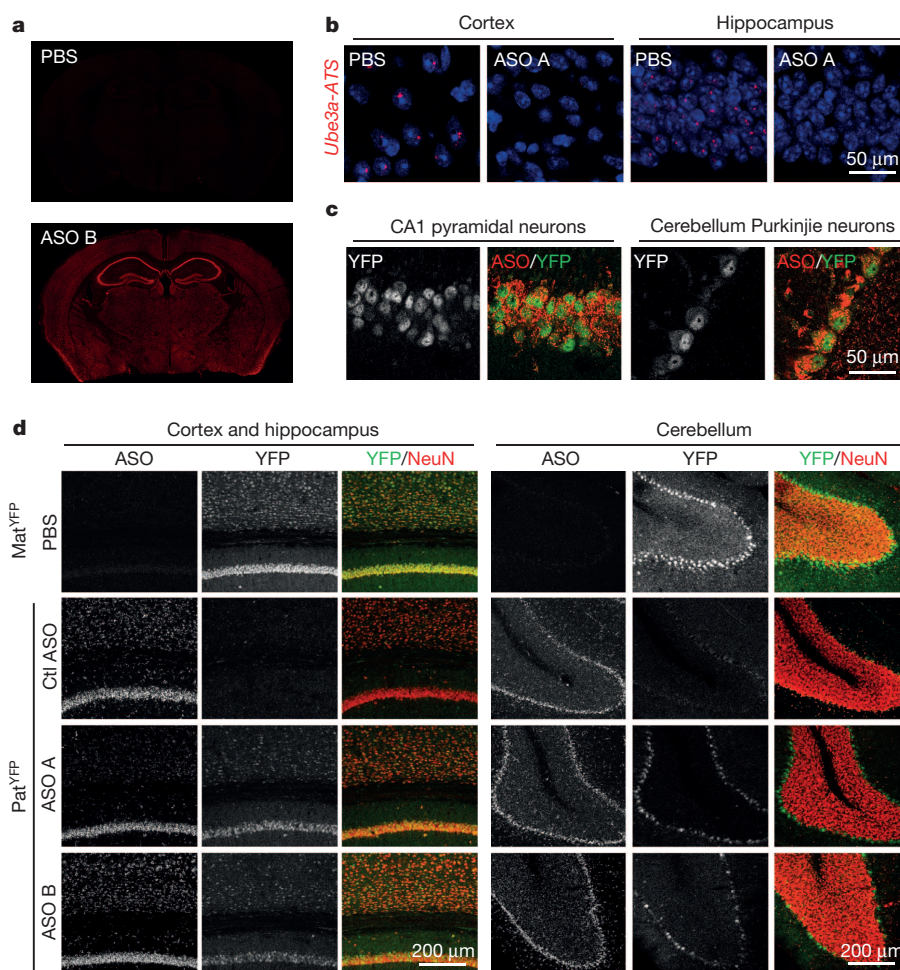


Figure 3 | Widespread distribution of paternal UBE3A unsilencing throughout the brain. **a–d**, Imaging of brain sections 4 weeks after treatment with Ctl ASO or ASO A and/or B in Pat^{YFP} mice, as labelled. **a**, ASO immunofluorescence on whole brain coronal sections. **b**, *In situ* hybridization

of *Ube3a-ATS*. **c**, High-magnification staining of UBE3A–YFP and ASO. **d**, Immunofluorescence for ASO, UBE3A–YFP and NeuN in critical brain regions. Mat^{YFP} mice treated with PBS were included for expression comparison.

However, paternal unsilencing was not complete compared with the maternal UBE3A–YFP level, consistent with the western blot analysis. To further increase the concentration of ASO in the brain, intrahippocampal delivery of ASO A was performed in Pat^{YFP} mice and complete unsilencing of UBE3A–YFP was observed near the injection site (Extended Data Fig. 8).

On the basis of the ability of ASO A to upregulate UBE3A, it was chosen for assessment of phenotypic correction in Angelman syndrome mice. Angelman syndrome mice phenocopy the impaired motor coordination and memory deficit observed in patients with the disease¹⁵. They have additional phenotypes including obesity, hypoactivity and decreased marble burying behaviour^{18–20}. Sex-matched Angelman syndrome littermates at 2–4 months of age were treated with ASO A or non-targeting control ASO (Ctl ASO). To determine the ability of ASO A to correct expression and behaviours relative to wild-type levels, a group of PBS-treated wild-type mice was included. After a single ICV injection, Angelman syndrome mice treated with ASO A showed reduction of *Ube3a-ATS* and partial restoration of UBE3A protein in the cortex ($35 \pm 19\%$), hippocampus ($35 \pm 15\%$) and cerebellum ($47 \pm 7\%$) compared with wild-type mice (Fig. 4b and Extended Data Fig. 9). UBE3A immunofluorescence also showed partial restoration of UBE3A protein in these brain regions (Fig. 4c and Extended Data Fig. 9). Four weeks after treatment, the mice were subjected to behavioural tests. A reversal of contextual freezing comparable to normal behaviour was observed in ASO-A-treated Angelman syndrome mice (analysis of variance (ANOVA), $F(2,39) = 5.242$, $P < 0.01$), indicating that the memory impairment was reversed

(Fig. 4d and Extended Data Fig. 9). However, there was no difference between mice treated with ASO A or Ctl ASO in open field, marble burying and accelerating rotarod tests (Extended Data Fig. 9). Complete phenotypic reversal may require treatment before a critical developmental window, a longer recovery time for rewiring of neural circuits, or a higher UBE3A induction level. Body weight was measured in a set of female mice that were injected at 3 months of age and followed for 5 months (Fig. 4e and Extended Data Fig. 9). The obesity phenotype in Angelman syndrome mice was corrected 1 month after treatment, and body weight remained significantly decreased compared with control ASO-treated mice for 5 months.

The genomic organization and regulation at the imprinting control centre is highly conserved between mouse and human. Therefore, ASO-mediated reduction of *UBE3A-ATS* is expected to restore UBE3A messenger RNA and protein in Angelman syndrome patient neurons. It is believed that maternal deficiency of UBE3A causes the majority of phenotypic findings in Angelman syndrome, and it is reasonable to expect that all patients with the condition, regardless of exact genotype, would benefit enormously from restored UBE3A expression. ASO therapy has been tested for neurological diseases in non-human primates and human clinical trials via intrathecal administration, with no serious adverse events^{16,21–23}. The well-tolerated delivery, broad tissue distribution and long duration of action indicate that ASOs may be a viable therapeutic strategy for CNS diseases and highlights the potential of an ASO drug for Angelman syndrome.

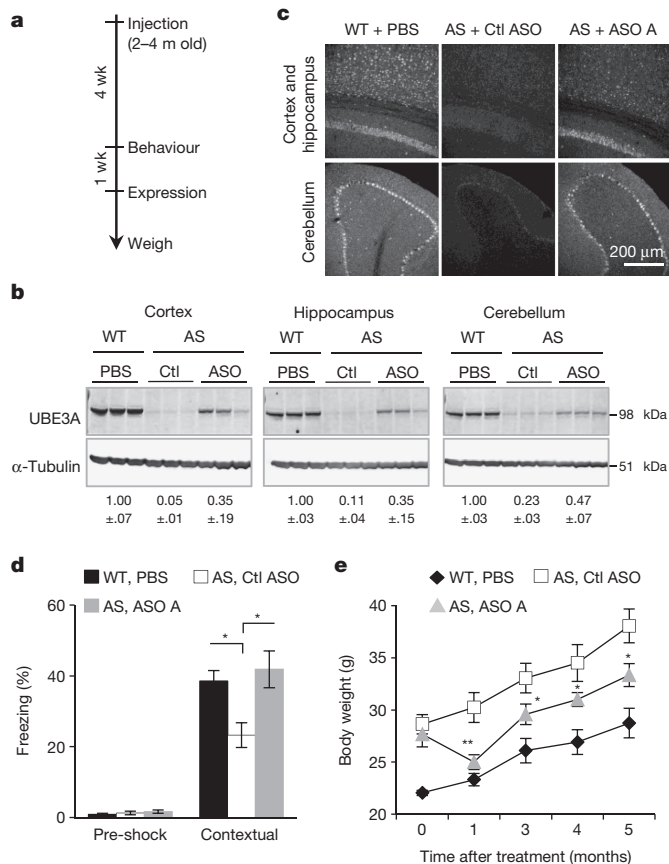


Figure 4 | ASO administration in adult Angelman syndrome mice unsilences paternal UBE3A and ameliorated abnormal phenotypes.

a, Experimental schedule. m, months; wk, weeks. **b**, Western blot with anti-UBE3A in brain regions of treated mice. Quantification of UBE3A normalized to α -tubulin is indicated below the images. Ctl, Ctl ASO. **c**, UBE3A immunofluorescence in wild-type (WT) or Angelman syndrome (AS) mice. **d**, Contextual fear measured during the fear conditioning assay. * $P < 0.05$, one-way analysis of variance (ANOVA) with Newman–Keuls post-hoc, $n = 13$ –15 per group. **e**, Growth curve of age-matched female mice. * $P < 0.05$, ** $P < 0.01$ (ASO A versus Ctl ASO), two-way ANOVA of repeated measurements with Newman–Keuls post-hoc, $n = 5$ per group.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 9 April; accepted 15 October 2014.

Published online 1 December 2014.

- Dagli, A., Buiting, K. & Williams, C. A. Molecular and clinical aspects of Angelman syndrome. *Mol. Syndromol.* **2**, 100–112 (2012).
- Williams, C. A., Driscoll, D. J. & Dagli, A. I. Clinical and genetic aspects of Angelman syndrome. *Genet. Med.* **12**, 385–395 (2010).
- Kishino, T., Lalande, M. & Wagstaff, J. UBE3A/E6-AP mutations cause Angelman syndrome. *Nature Genet.* **15**, 70–73 (1997).

- Matsuura, T. *et al.* De novo truncating mutations in E6-AP ubiquitin-protein ligase gene (UBE3A) in Angelman syndrome. *Nature Genet.* **15**, 74–77 (1997).
- Albrecht, U. *et al.* Imprinted expression of the murine Angelman syndrome gene, *Ube3a*, in hippocampal and Purkinje neurons. *Nature Genet.* **17**, 75–78 (1997).
- Rougeulle, C., Cardoso, C., Fontes, M., Colleaux, L. & Lalande, M. An imprinted antisense RNA overlaps UBE3A and a second maternally expressed transcript. *Nature Genet.* **19**, 15–16 (1998).
- Chamberlain, S. J. & Brannan, C. I. The Prader–Willi syndrome imprinting center activates the paternally expressed murine *Ube3a* antisense transcript but represses paternal *Ube3a*. *Genomics* **73**, 316–322 (2001).
- Meng, L., Person, R. E. & Beaudet, A. L. *Ube3a*-ATS is an atypical RNA polymerase II transcript that represses the paternal expression of *Ube3a*. *Hum. Mol. Genet.* **21**, 3001–3012 (2012).
- Meng, L. *et al.* Truncation of *Ube3a*-ATS unsilences paternal *Ube3a* and ameliorates behavioral defects in the Angelman syndrome mouse model. *PLoS Genet.* **9**, e1004039 (2013).
- Huang, H. S. *et al.* Topoisomerase inhibitors unsilence the dormant allele of *Ube3a* in neurons. *Nature* **481**, 185–189 (2012).
- Wu, H. *et al.* Determination of the role of the human RNase H1 in the pharmacology of DNA-like antisense drugs. *J. Biol. Chem.* **279**, 17181–17189 (2004).
- Dindot, S. V., Antalffy, B. A., Bhattacharjee, M. B. & Beaudet, A. L. The Angelman syndrome ubiquitin ligase localizes to the synapse and nucleus, and maternal deficiency results in abnormal dendritic spine morphology. *Hum. Mol. Genet.* **17**, 111–118 (2008).
- Cassidy, S. B., Schwartz, S., Miller, J. L. & Driscoll, D. J. Prader–Willi syndrome. *Genet. Med.* **14**, 10–26 (2012).
- King, I. F. *et al.* Topoisomerases facilitate transcription of long genes linked to autism. *Nature* **501**, 58–62 (2013).
- Jiang, Y. H. *et al.* Mutation of the Angelman ubiquitin ligase in mice causes increased cytoplasmic p53 and deficits of contextual learning and long-term potentiation. *Neuron* **21**, 799–811 (1998).
- Kordasiewicz, H. B. *et al.* Sustained therapeutic reversal of Huntington's disease by transient repression of huntingtin synthesis. *Neuron* **74**, 1031–1044 (2012).
- Rigo, F. *et al.* Pharmacology of a central nervous system delivered 2'-O-methoxyethyl-modified survival of motor neuron splicing oligonucleotide in mice and nonhuman primates. *J. Pharmacol. Exp. Ther.* **350**, 46–55 (2014).
- Cattanach, B. M. *et al.* A candidate model for Angelman syndrome in the mouse. *Mamm. Genome* **8**, 472–478 (1997).
- Allensworth, M., Saha, A., Reiter, L. T. & Heck, D. H. Normal social seeking behavior, hypoactivity and reduced exploratory range in a mouse model of Angelman syndrome. *BMC Genet.* **12**, 7 (2011).
- Huang, H. S. *et al.* Behavioral deficits in an Angelman syndrome model: effects of genetic background and age. *Behav. Brain Res.* **243**, 79–90 (2013).
- Smith, R. A. *et al.* Antisense oligonucleotide therapy for neurodegenerative disease. *J. Clin. Invest.* **116**, 2290–2296 (2006).
- Miller, T. M. *et al.* An antisense oligonucleotide against *SOD1* delivered intrathecally for patients with *SOD1* familial amyotrophic lateral sclerosis: a phase 1, randomised, first-in-man study. *Lancet Neurol.* **12**, 435–442 (2013).
- Chiriboga, C. *et al.* Results of an open-label, escalating dose study to assess the safety, tolerability, and dose range finding of a single intrathecal dose of ISIS-SMNRx in patients with spinal muscular atrophy. 65th American Academy of Neurology Annual Meeting Abstract S36.002 (2013).

Acknowledgements We thank T. Cooper for suggesting the collaboration between L.M., A.L.B. and Isis Pharmaceuticals. L.M. thanks D. Liu, X. Jun, H. Zheng, J. Rosen, C. Spencer and X. Zhai for technical support and reagent sharing, and M. Costa-Mattioli, J. Jankowsky and Y. Elgersma for discussions. This research was supported by funding from National Institutes of Health grant R01 HD037283 (A.L.B.), the Angelman Syndrome Foundation (A.L.B. and L.M.), and Baylor College of Medicine Intellectual and Developmental Disability Research Center grant P30HD024064.

Author Contributions L.M. and A.J.W. designed and performed experiments, analysed data, and wrote the paper (equal contribution). S.C. performed ASO delivery for Figure 2. C.F.B., A.L.B. and F.R. supervised the project. All authors discussed the experimental results.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Readers are welcome to comment on the online version of the paper. The authors declare competing financial interests: details are available in the online version of the paper. Correspondence and requests for materials should be addressed to A.L.B. (abeaudet@bcm.edu) or F.R. (frigo@isisph.com).

METHODS

Animals. All the animals of wild-type, *Ube3a*^{KO/+} (ref. 15), and *Ube3a*^{+/-YFP} (ref. 12) genotypes were kept on C57BL/6 background and housed under standard conditions in a pathogen-free mouse facility. All animal procedures were performed in accordance with National Institutes of Health guidelines and approved by the Institutional Animal Care and Use Committee at Baylor College of Medicine and Isis Pharmaceuticals. To generate Angelman syndrome mice, *Ube3a*^{KO/+} mice were born to mothers who carry the mutation on their paternal chromosome. Wild-type and Angelman syndrome littermates were housed in the same cage whenever possible.

Oligonucleotide synthesis. Synthesis and purification of all chemically modified oligonucleotides was performed as previously described²⁴. The 2'-MOE gapmer ASOs are 20 nucleotides in length, wherein the central gap segment comprising ten 2'-deoxynucleotides is flanked on the 5' and 3' wings by five 2'-MOE modified nucleotides. Internucleoside linkages are purely phosphorothioate (ASO B) or interspersed with phosphodiester (ASO A), and all cytosine residues are 5'-methylcytosines. The RNase H inactive ASO consists of 20 2'-MOE modified nucleotides. The sequences of the ASOs are as follows: Ctl ASO (*in vitro*), 5'-CCTTCCCTGAAGGTTCTCTCC-3'; Ctl ASO (*in vivo*), 5'-CTCAGTAACATTGACACCAC-3' (ref. 25); ASO A, 5'-GATCCATTTGTGTTAAGCTG-3'; ASO B, 5'-CCAGCCTTGTGGATATCA T-3'; ASO116, 5'-CAGAGTTTTCACTCATTTTG-3'.

Primary neuron culture and ASO treatment. Primary cultures of hippocampal and cortical neurons were established as previously described⁸ from P0–P2 offspring of wild-type C57BL/6 or *Ube3a*^{+/-YFP} mice. Four days after plating, half of the medium was replaced and the ASO (10 μ M) or topotecan (300 nM) was added to the culture medium for 72 h, unless otherwise noted. Arabinofuranosyl cytidine (Sigma) was used to inhibit glial proliferation.

Immunofluorescence. Primary neurons were fixed with 4% paraformaldehyde (PFA) for 1 h and washed in PBS. For *in vivo* samples, mice were anaesthetized and perfused with PBS and 4% PFA. Brain tissue was fixed with PFA overnight and dehydrated in 30% sucrose. Coronal sections of 35 μ m were prepared and stained as previously described⁹. The following antibodies were used: anti-GFP (ab13970, Abcam, 1:1,000), anti-NeuN (MAB377, Millipore, 1:1,000 dilution), anti-UBE3A (A300-352A, Bethyl Laboratories, 1:500), and anti-ASO (Isis, 1:10,000)²⁶. For the high-throughput *in vitro* ASO screen, the plates were imaged with ImageXpress^{Ultra} confocal system (Molecular Device) and then further processed with the MetaXpress software (Molecular Device). Typically 200–800 cells were scored per well and the signal intensities were averaged and normalized to untreated control cells. For tissue sections, images were taken using a confocal microscope (Leica).

qRT-PCR. Total cellular RNA was isolated from cultured neurons and mouse tissue using the RNeasy kit (Qiagen). For preparation of mouse tissue, samples were first lysed using FastPrep Lysing Matrix Tubes (MP-Biomedicals) in RLT buffer containing 1% β -mercaptoethanol. On-column DNase digestion was performed for all samples. For qRT-PCR, approximately 10 ng RNA was added to EXPRESS One-Step SuperScript qRT-PCR Kit (Life Technologies) with Taqman primer and probe sets or EXPRESS One-Step SYBR GreenER Kit (Life Technologies) with SYBR primer sets (see Extended Data Table 1 for sequences). All quantification was performed by the relative standard curve method and normalized to total RNA by Ribogreen or to the housekeeping genes *Gapdh*.

DNA methylation analysis. Primary neuron cultures were derived from the F1 hybrid of CAST.chr7 male and C57BL/6 female mice and treated with ASO (10 μ M, 72 h). Genomic DNA was then extracted and processed for bisulphite sequencing of the PWS imprinting centre at the *Snrpn* DMR1 region (*Snrpn* promoter and exon 1) as previously described⁸.

Northern blot. Total RNA was isolated from ASO-treated primary neurons (10 μ M, 72 h) by TRIzol (Life Technologies) according to the manufacturer's protocol. Three micrograms total RNA was separated on an 8% polyacrylamide-7M urea gel, and then transferred by semi-dry transfer (12 V, 30 min) to GeneScreen plus hybridization transfer membrane (Perkin Elmer). The northern probes were 5'-end labelled with ATP Gamma-³²P (Perkin Elmer) using T4 polynucleotide kinase (New England Biolabs), and then hybridized to the membrane at 42 °C for 30 min. After washing membrane in wash buffer (2 \times SSC containing 0.1% SDS), the membrane was exposed to a PhosphorImager and quantified. The oligonucleotide probe sequences used were *Snord116* 5'-TTCCGATGAGAGTGGCGGTACAGA-3' and 5.8S rRNA 5'-TCCTGCAATTCACATTAATTCTCGCAGCTAGC-3'.

Western blot. Cultured neurons and mouse tissue were homogenized and lysed in RIPA buffer (Sigma-Aldrich) containing EDTA-free cOmplete Protease Inhibitor Cocktail (Roche). Protein concentration of the supernatant was determined by the DC protein assay (Bio-Rad). Ten to forty micrograms of protein lysate was separated on a precast 4–20% Bis-Tris gel (Life Technologies) and transferred by iBlot (Life Technologies). The following primary antibodies were diluted in Odyssey blocking buffer: anti-UBE3A (611416, BD Biosciences, 1:500), anti-GFP (NB600-308, Novus Biologicals, 1:500), anti- β -tubulin (T9026, Sigma, 1:20,000) and anti- α -tubulin

(T5168, Sigma, 1:8,000). Following primary antibody incubation, membranes were probed with goat anti-rabbit IRDye 680LT (LiCor) or goat anti-mouse IRDye 800CW (LiCor) and imaged and quantified using the LiCor Odyssey system.

ASO *in vivo* administration. Lyophilized ASOs were dissolved in sterile PBS without calcium or magnesium and quantified by ultraviolet spectrometry. The ASOs were then diluted to the desired concentration required for dosing mice and sterilized through a 0.2 μ m filter. Mice were anaesthetized with 2% isoflurane and placed in a stereotaxic frame (David Kopf Instruments). After exposing the skull, a needle (Hamilton, 1701 RN 10 μ l micro syringe, needle 26 s/2''/2) was used to penetrate the skull at 0.2 mm posterior and 1.0 mm lateral to the bregma, and lowered to a depth of 3.0 mm, to deliver PBS or ASO (ASO A, 700 μ g; ASO B, 500 μ g) at a rate of approximately 1 μ l per 30 s. The needle was left in place for 5 min, slowly withdrawn and the incision was sutured. For intrahippocampal injection, the coordinate of \sim 2.0 mm anterior, 1.5 mm lateral and \sim 2.0 mm dorsal to the bregma was used.

Fluorescence *in situ* hybridization. Tissue preparation and RNA fluorescence *in situ* hybridization (FISH) was carried out by the RNA In-Situ Hybridization Core at Baylor College of Medicine, as previously described⁹. Primers for DNA template synthesis are 5'-ATTTAGGTGACACTATAGAAGCGAAGATGAGTCAGTTTGGTTTT-3' and 5'-TAATACGACTCACTATAGGGAGATTCTGAGTCTTCTTCCATAGC-3'. The T7 promoter was used to generate the *Ube3a*-ATS probe.

Behavioural tests. Three groups of age- and sex-matched littermates were generated, and mice were randomly assigned to each treatment group. At 2 to 4 months of age, Angelman syndrome mice received a single 700 μ g dose of non-targeting control ASO or ASO A. Wild-type mice injected with an equal volume of PBS were included as controls. Four weeks after treatment, a battery of behavioural tests was performed by an experimenter blind to the genotype and treatment group using a protocol previously described⁹ in the Neurobehavioural Core at Baylor College of Medicine. The open field and marble burying tests were performed on day 1, the accelerating rotarod test was performed on day 2 and 3, and the fear conditioning test was performed on day 4 and 5. Mice were acclimated to the test room for 30 min before each behaviour test.

For the open field assay each mouse was placed in the centre of a clear Plexiglas (40 \times 40 \times 30 cm) open-field arena (Versamax Animal Activity Monitor, AccuScan Instruments) and allowed 30 min to explore. Overhead lighting was \sim 800 lx inside the field, and the white noise was at \sim 60 dB. Mouse activity was recorded and quantified.

For the marble burying test, each mouse was placed in a standard mouse cage containing 20 small (1.5–2 cm) clean black marbles on top of 4 inches of corn cob bedding, forming 4 rows of 5 columns. After a period of 30 min exploration, the mouse was removed from the cage and the number of marbles buried at least 50% was recorded.

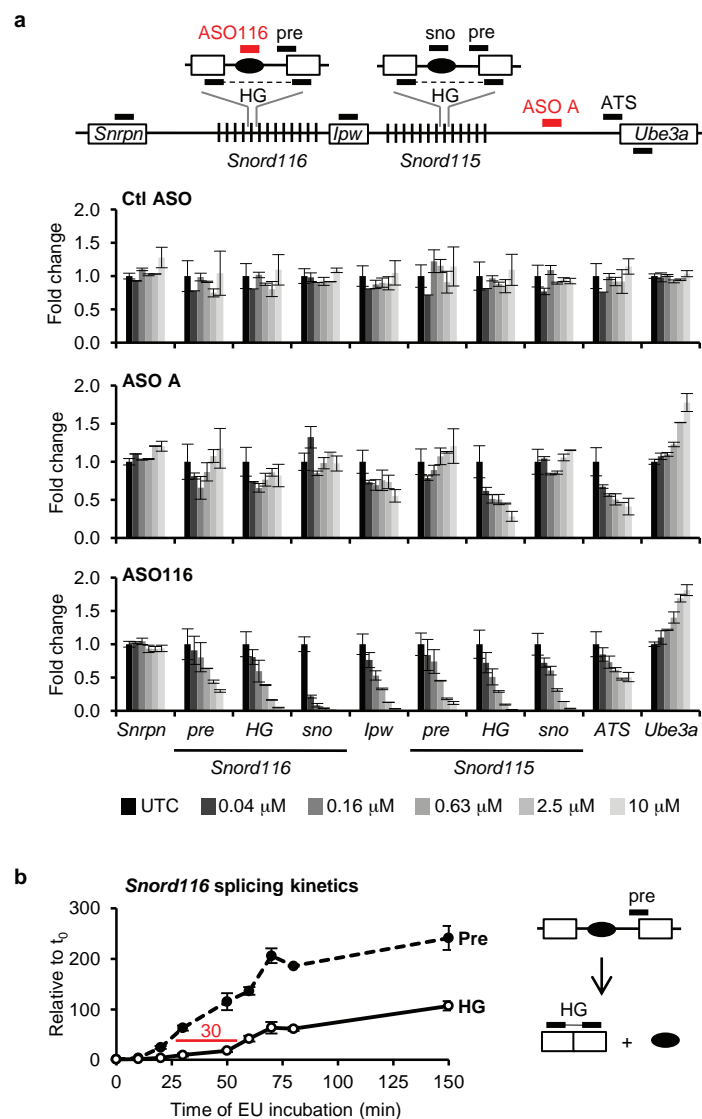
For the accelerating rotarod, the test was performed with a rotating rod system that rotates from 4 to 40 r.p.m. within 5 min (model 7650 Rota-rod, Ugo Basile). Mice were placed on the rotating rod and the time until falling off or losing balance (mice not walking on the rod for two consecutive turns) was recorded. For two consecutive days, four trials were performed per day with at least a 30 min interval between trials.

For contextual fear conditioning, on the training day, each mouse was placed in a test chamber. After 2 min of free exploration (baseline/pre-shock freezing), the mouse received an auditory tone (2,800 Hz, 85 db, 30 s) followed by a foot-shock (0.7 mA, 2 s). The training was repeated once. The mouse remained in the chamber for one additional min (post-shock freezing) and then was returned to the home cage. Twenty-four hours after training, mice were returned to the same test chamber for 5 min and tested for freezing in response to the training context (contextual freezing). Afterwards, the environmental settings of the test chamber were drastically altered and the mice were placed back in the modified context. They were allowed 3 min of free exploration, and then the auditory tone was presented for 3 min to test the fear response to the cue (cued freezing). Freezing frequency was analysed with FreezeFrame software (San Diego Instruments).

Isolation of nascent RNA. Nascent RNA was isolated using the Click-iT Nascent RNA Capture Kit (Life Technologies), according to the manufacturer's protocol. In brief, wild-type primary neurons were incubated with 5-ethynyl uridine (EU, 0.5 mM) for 0 to 150 min at which time total RNA was isolated by TRIzol. Five micrograms total RNA was biotinylated with 0.5 mM biotin azide, and 500 ng biotinylated RNA was precipitated on streptavidin beads. Nascent EU-containing RNA captured on the beads was used for SuperScript VILO cDNA synthesis (Life Technologies) followed by qPCR.

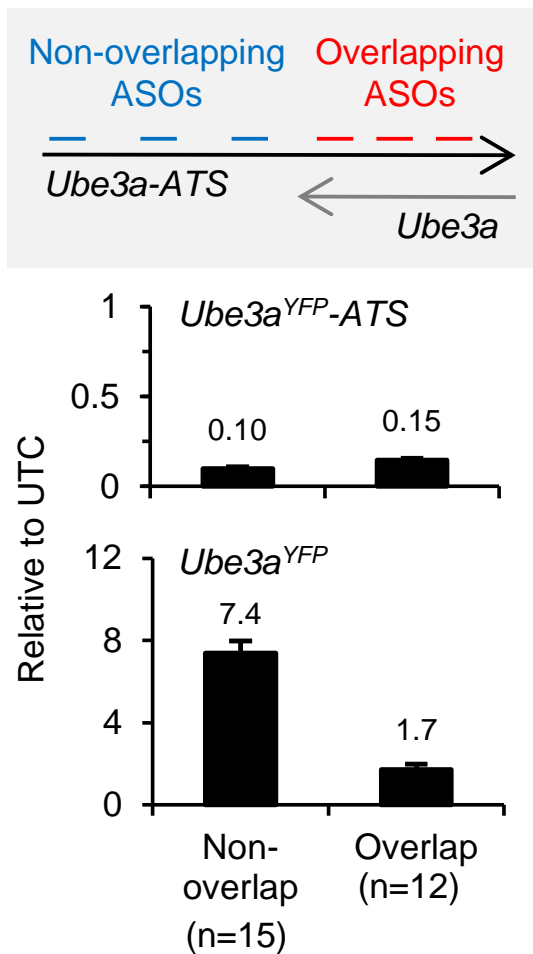
24. Swayze, E. E. *et al.* Antisense oligonucleotides containing locked nucleic acid improve potency but cause significant hepatotoxicity in animals. *Nucleic Acids Res.* **35**, 687–700 (2007).

25. Lagier-Tourenne, C. *et al.* Targeted degradation of sense and antisense C9orf72 RNA foci as therapy for ALS and frontotemporal degeneration. *Proc. Natl Acad. Sci. USA* **110**, E4530–E4539 (2013).
26. Butler, M., Stecker, K. & Bennett, C. F. Cellular distribution of phosphorothioate oligodeoxynucleotides in normal rodent tissues. *Lab. Invest.* **77**, 379–388 (1997).



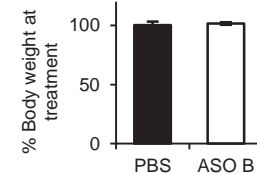
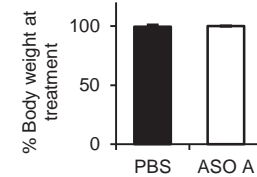
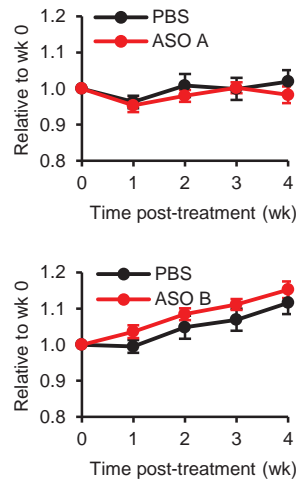
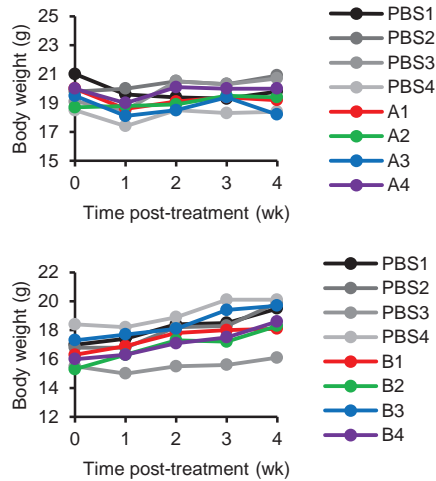
Extended Data Figure 1 | ASOs targeting *Snord116* reduced *Ube3a-ATS* pre-mRNA. **a**, Top, schematic of the ASO-binding sites and location of qRT-PCR primer and probe sets. Bottom, qRT-PCR from wild-type primary neurons treated with ASO A or ASO116 (72 h) using primer and probe sets to the indicated regions of *Ube3a-ATS* pre-mRNA and mRNA. **b**, Nascent transcripts were isolated from wild-type primary neurons incubated with 5-ethynyl uridine (see Methods) for the indicated time. qRT-PCR for

pre-mRNA and mature mRNA within the *Snord116* region. The red line indicates the 30 min delay between the appearance of pre-mRNA and mature mRNA. Assuming a transcription elongation rate of 4 kb min^{-1} , it would take RNAPII 80 min to transcribe the 332 kb distance from the last copy of *Snord116* to the ASO-binding site. HG, host gene. $n = 2$ per group, mean \pm absolute deviation.

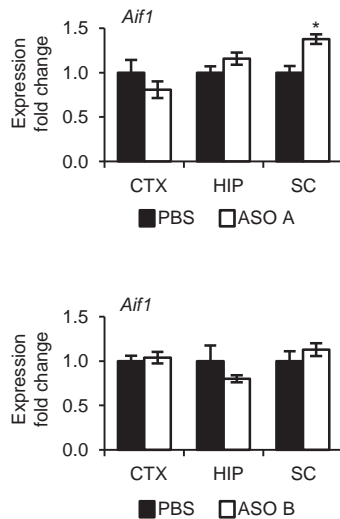


Extended Data Figure 2 | ASOs complementary to two regions of *Ube3a-ATS* differed in their ability to unsilence paternal *Ube3a*. Pat^{YFP} primary neurons were treated with ASOs that bind *Ube3a-ATS* 5' of *Ube3a* (non-overlap ASOs, $n = 15$) or that bind to the gene body region (overlap ASOs, $n = 12$) for 72 h. The level of *Ube3a^{YFP}-ATS* reduction and *Ube3a^{YFP}* upregulation was analysed by qRT-PCR and normalized to untreated control (UTC) neurons. Data are shown as mean \pm s.e.m.

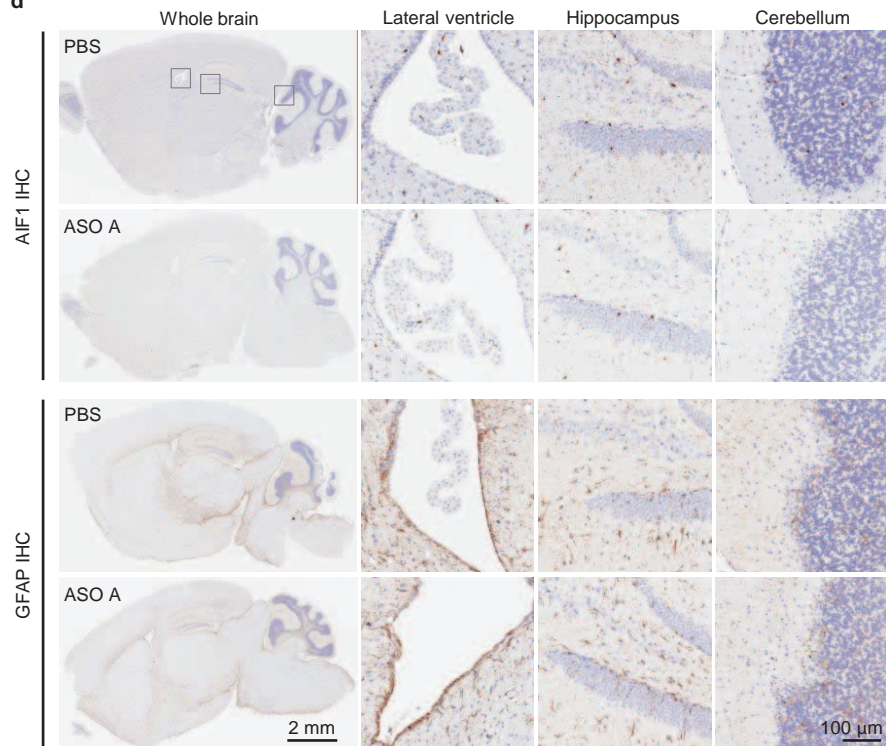
a



c



d

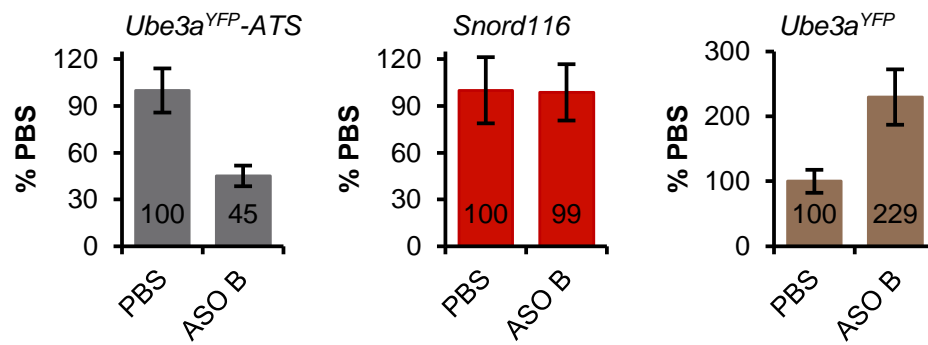


Extended Data Figure 3 | *In vivo* ASO administration was well tolerated.

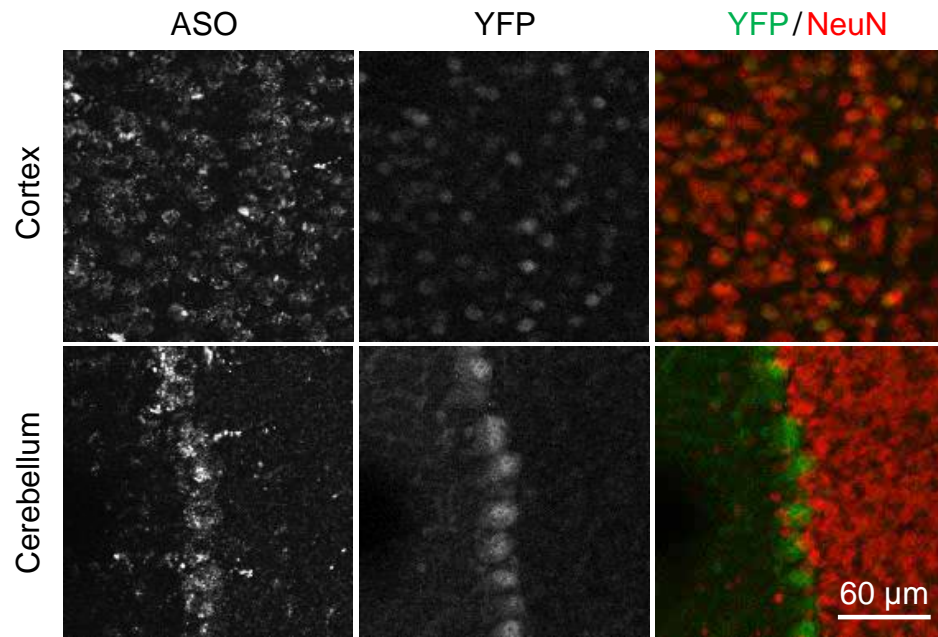
a, Left, body weight of individual wild-type C57BL/6 female mice (2 months old) treated with PBS or ASO measured weekly for 4 weeks after treatment. Right, change in body weight at each time point relative to body weight at time of treatment. $n = 4$ per group, mean \pm s.e.m. **b**, Per cent change in body weight of Pat^{YFP} mice 4 weeks after treatment relative to pre-treatment.

$n = 3-4$, mean \pm s.e.m. **c**, Microglial activation was measured by *Aif1* qRT-PCR 4 weeks after treatment. CTX, cortex; HIP, hippocampus; SC, thoracic spinal cord. $*P < 0.05$, two-tailed t -test, $n = 3-4$ per group, mean \pm s.e.m. **d**, Immunohistochemistry for AIF1 and GFAP on sagittal brain sections from wild-type C57BL/6 female mice treated with PBS or ASO for 2 weeks.

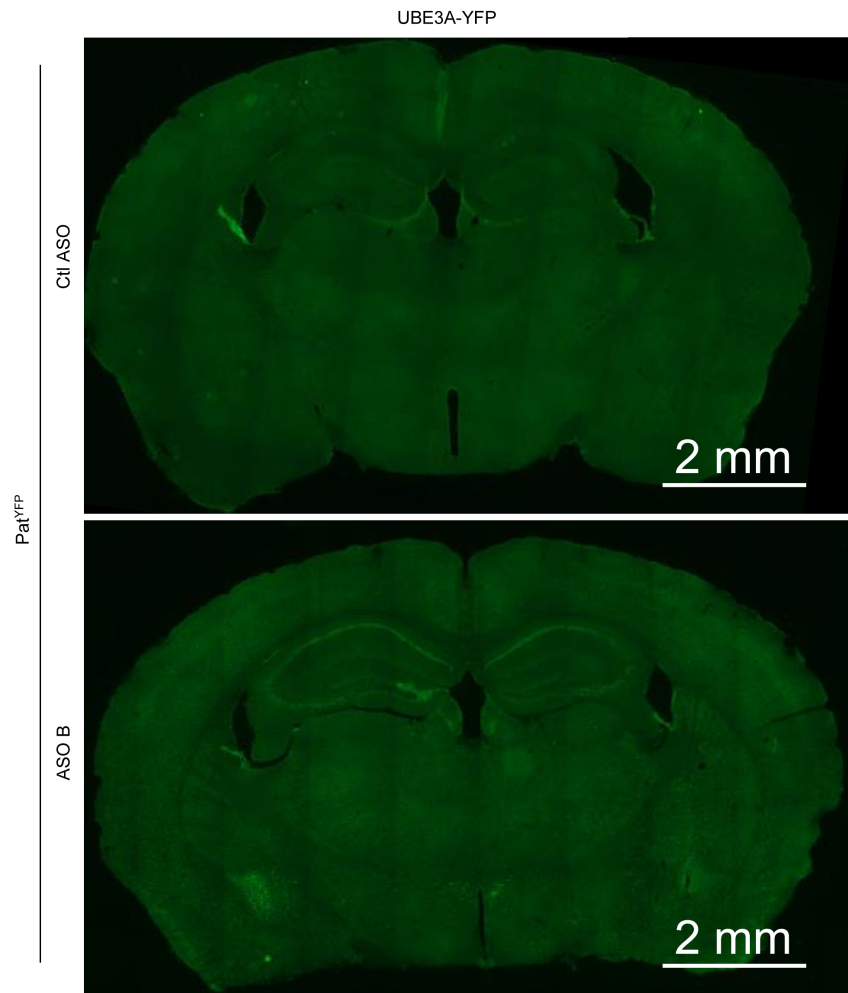
Hypothalamus



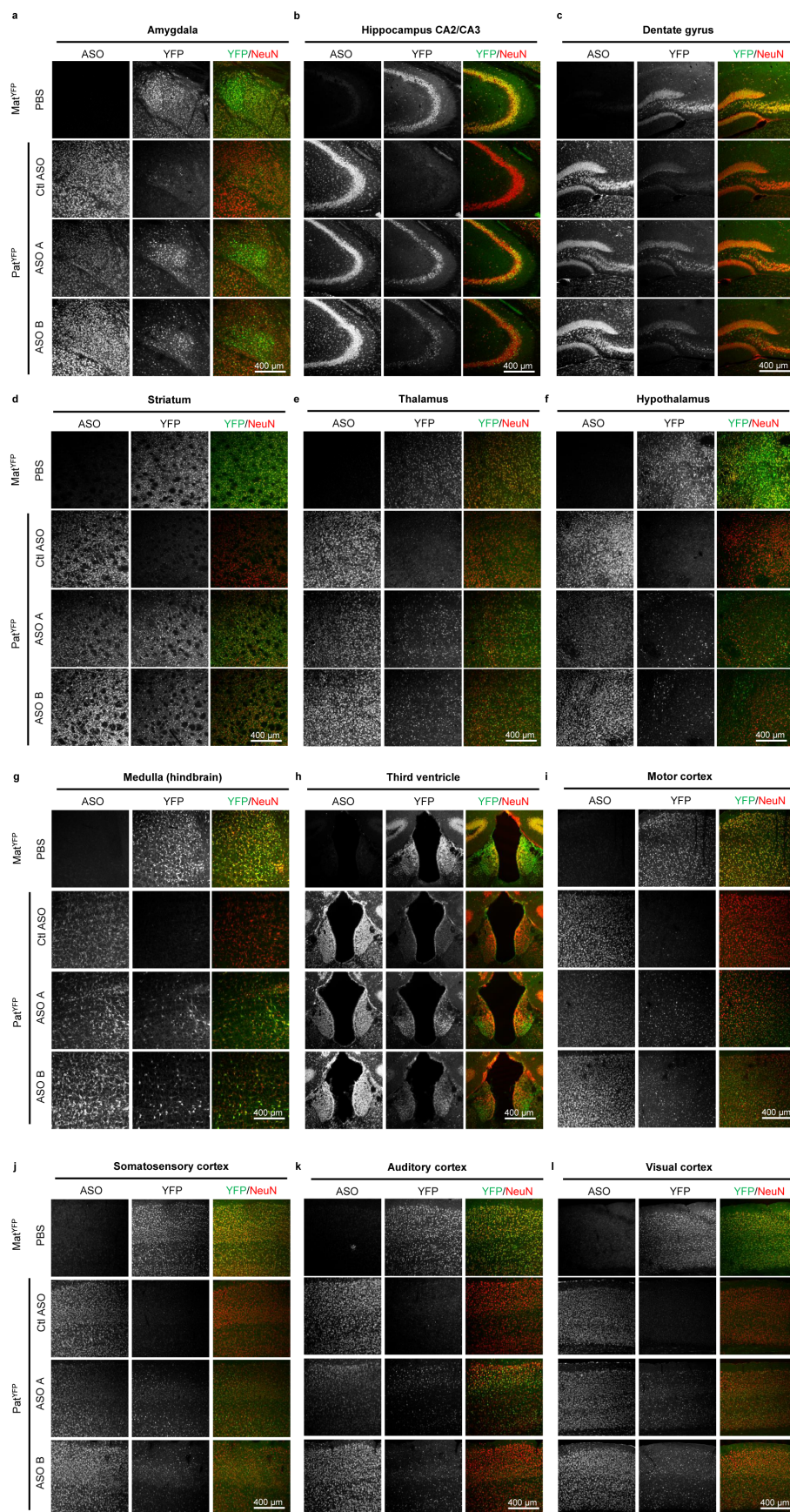
Extended Data Figure 4 | *Snord116* was not reduced in the hypothalamus. qRT-PCR on RNA isolated from *Pat^{YFP}* mice 4 weeks after treatment with PBS or ASO B.



Extended Data Figure 5 | UBE3A unsilencing persisted 4 months after treatment. ASO and YFP immunofluorescence on brain sections of cortex and cerebellum in Pat^{YFP} mice 4 months after treatment with ASO A.

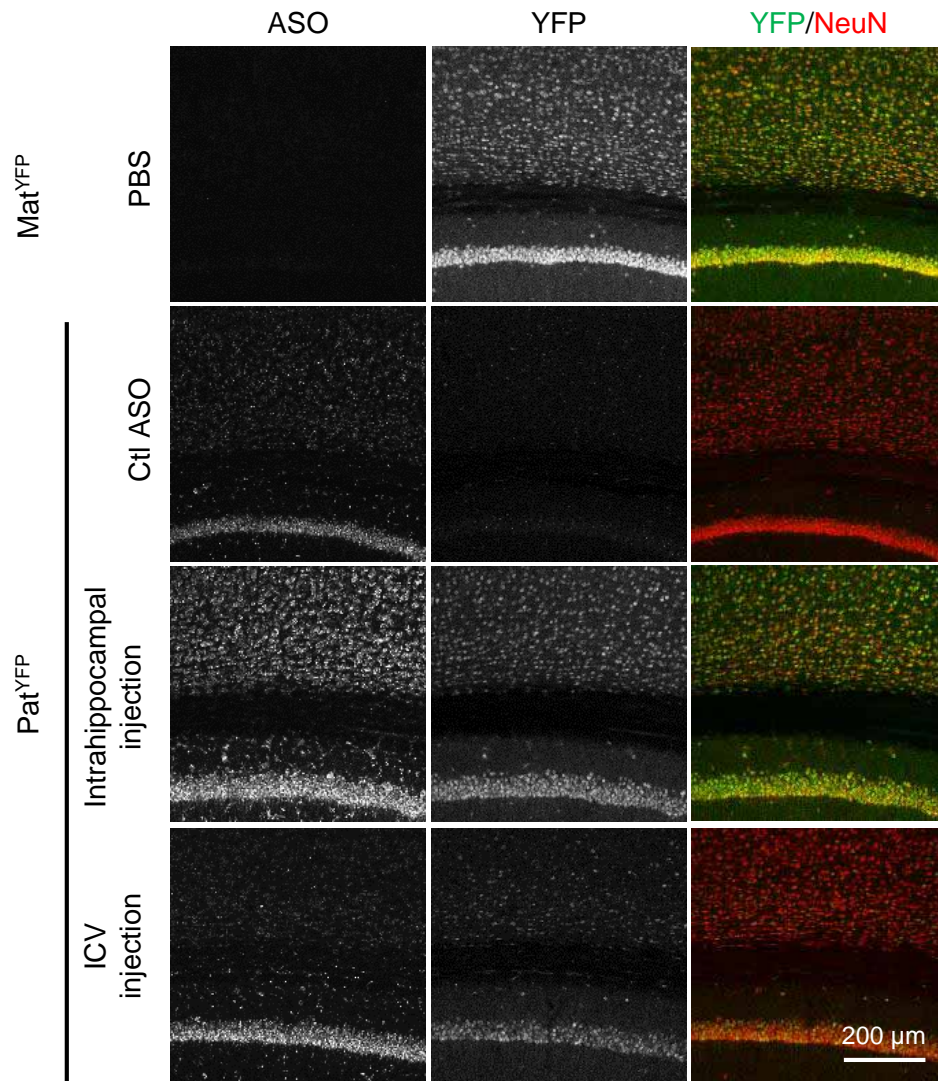


Extended Data Figure 6 | UBE3A–YFP was upregulated throughout the brain. Whole-brain image of YFP fluorescence in Pat^{YFP} mice treated with PBS or ASO 4 weeks after treatment.



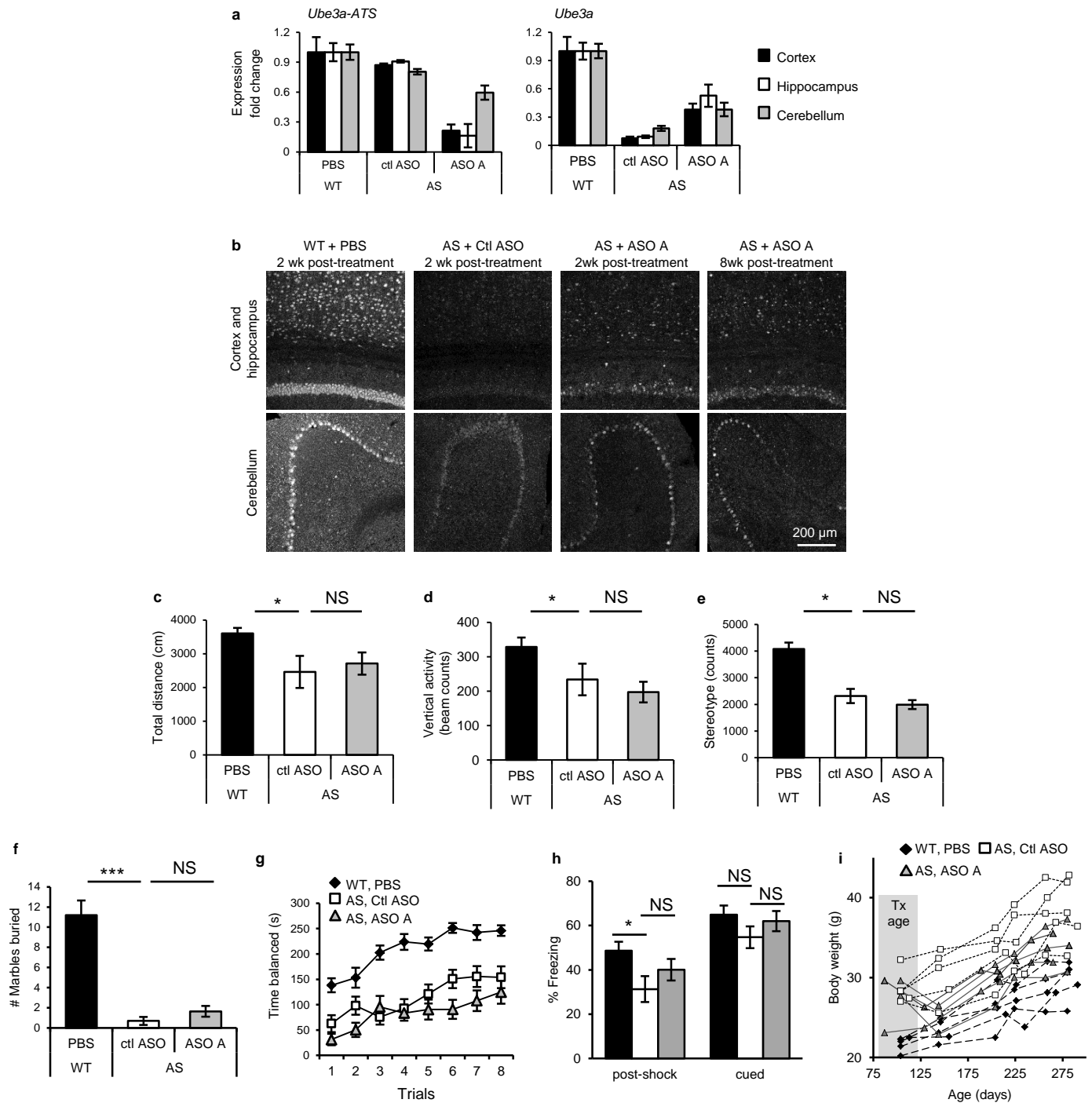
Extended Data Figure 7 | Imaging of unsilenced UBE3A-YFP in specific brain regions. a–l, Immunofluorescence for ASO, UBE3A-YFP and NeuN 4 weeks after treatment in Mat^{YFP} or Pat^{YFP} mice of the amygdala

(a), hippocampus CA2 and CA3 layers (b), dentate gyrus (c), striatum (d), thalamus (e), hypothalamus (f), medulla (g), third ventricle (h), motor cortex (i), somatosensory cortex (j), auditory cortex (k) and visual cortex (l).



Extended Data Figure 8 | Intrahippocampal injection of ASO A in Pat^{YFP} mice resulted in near complete unsilencing of paternal UBE3A–YFP. YFP immunofluorescence on brain sections from Pat^{YFP} mice treated with Ctl

ASO, 100 μg ASO A via intrahippocampal injection, or 700 μg ASO A via ICV injection. A Mat^{YFP} mouse treated with PBS was included for comparison.



Extended Data Figure 9 | ASO treatment in Angelman syndrome mice upregulated *Ube3a*. **a**, RNA levels of *Ube3a-ATS* and *Ube3a* were determined by qRT-PCR in wild-type (WT) mice treated with PBS and Angelman syndrome (AS) mice treated with Ctl ASO or ASO A. $n = 2-3$ per group, mean \pm s.e.m. **b**, UBE3A immunofluorescence on brain sections was performed 2 to 8 weeks after treatment. **c-h**, ASO treatment in adult Angelman syndrome mice did not reverse some disease-associated phenotypes. **c**, Total distance travelled in the open field assay. **d**, Vertical activity in the open

field assay. **e**, Stereotype activity in the open field assay. **f**, Marble burying test. The y axis represents the number of marbles at least 50% buried. **g**, Accelerating rotarod test during eight trials. **h**, Post-shock and cued response measured during the fear conditioning assay. $n = 13-15$ per group. * $P < 0.05$, *** $P < 0.001$ (one-way ANOVA with Newman-Keuls post-hoc analysis). NS, not significant. **i**, Growth curve of age-matched female mice. Each line represents weight measurements of a single mouse over a 5-month time course post-injection, $n = 5$ per group. Tx age, age of mouse at time of treatment.

Extended Data Table 1 | qRT-PCR primer sequences

Target	Forward Sequence	Reverse Sequence	Probe Sequence (Taqman)
<i>Ube3a^{YFP}-ATS</i>	CCAATGACTCATGATTGTCCTG	GGTACCCGGGGATCCTCTAG	
<i>Ube3a^{YFP}</i>	TGGAGGACTAGGAAAATTGAAGATG	CGCCCTTGCTCACCATG	CCAAAAATGGCCCAGACACAGAAAGG
<i>Ube3a-ATS</i>	CCAATGACTCATGATTGTCCTG	GTGATGGCCTTCAACAATCTC	
<i>Ube3a</i>	GCACCTGTTGGAGGACTAGG	GTGATGGCCTTCAACAATCTC	
<i>Snrpn</i>	TGTGATTGTGATGAGTTCAGGAAGA	ACCAGACCCAAAACCCGTTT	CAAGCCAAAGAATGCAAAACAGCCAGAA
<i>Snord116</i>	GGATCTATGATGATTCCCAG	GGACCTCAGTTCCGATGA	
<i>Snord116HG</i>	TGTGCTGACTTGCCCTAG	GTTCGATGGAGACTCAGTTGG	AAACATGCAGAGGAAATGGCCCC
<i>Snord116pre</i>	ATTGGTCCCACTGTAATCGG	GTTCGATGGAGACTCAGTTGG	AAACATGCAGAGGAAATGGCCCC
<i>Snord115</i>	CTGGGTCAATGATGACAAC	TTGGGCCTCAGCGTAATCC	
<i>Snord115HG</i>	CAGCAATCCCTCTCCAGTTC	AAGGTGGCATGTGAGATGAC	TGTGACCATTCTACTCTGAGCCAGTT
<i>Snord115pre</i>	CCATGTGACCATTCTACTCTG	AGAATTCGGCTACATCTACTTGG	TGGAAAGGAAGGTAAGTGTGGATTAGGT
<i>lpw</i>	GCTGATAACATTCACTCCCAGA	GAATGAGCTGACAACCTACTCC	TTGGACACCCCTGCAGAAGATGACTT
<i>Gapdh</i>	GGCAAATTC AACGGCACAGT	GGGTCTCGCTCCTGGAAGAT	AAGGCCGAGAATGGGAAGCTTGTCATC
<i>Pgk1</i>	ATGTCGCTTTCCAACAAGCTG	GCTCCATTGTCCAAGCAGAAT	
<i>Nrxn3</i>	GATGAAGACTTTGTGAATGTGA	CCGTCTGATTCTGGCTCCGTG	GACCATCCCTGTTGTACTGC
<i>Astn2</i>	CAGCACCACTACAACCTCTCAC	TCACTCCTCCAGACGAAGTCACCA	CGCAGAATCAGATGAGCCTT
<i>Pchd15</i>	CGGCGAAGTCATTGGTGAA	ACCAACTTGATCATTCTTTTCTTGCCAC	GTTCTGCTTCTCTGCGACTC
<i>Csmd1</i>	GCTGCCATTCTTGTCCCT	ACTTTTGGTCTTGTCTGTGTTGTAGAGGT	TCAAATGAAGCTTGTCCATTACTG
<i>Il1rapl1</i>	CTTGAAATCCTCCCTGATATGCT	CCAACTGGAACATACATTGAAGATGTGGCT	CCGCTTACTTTGATCTACGCA
<i>Aif1</i>	TGGTCCCCCAGCCAAGA	CCCACCGTGTGACATCCA	AGCTATCTCCGAGCTGCCCTGATTGG

Intracellular α -ketoglutarate maintains the pluripotency of embryonic stem cells

Bryce W. Carey^{1*}, Lydia W. S. Finley^{2*}, Justin R. Cross³, C. David Allis¹ & Craig B. Thompson²

The role of cellular metabolism in regulating cell proliferation and differentiation remains poorly understood¹. For example, most mammalian cells cannot proliferate without exogenous glutamine supplementation even though glutamine is a non-essential amino acid^{1,2}. Here we show that mouse embryonic stem (ES) cells grown under conditions that maintain naive pluripotency³ are capable of proliferation in the absence of exogenous glutamine. Despite this, ES cells consume high levels of exogenous glutamine when the metabolite is available. In comparison to more differentiated cells, naive ES cells utilize both glucose and glutamine catabolism to maintain a high level of intracellular α -ketoglutarate (α KG). Consequently, naive ES cells exhibit an elevated α KG to succinate ratio that promotes histone/DNA demethylation and maintains pluripotency. Direct manipulation of the intracellular α KG/succinate ratio is sufficient to regulate multiple chromatin modifications, including H3K27me3 and ten-eleven translocation (Tet)-dependent DNA demethylation, which contribute to the regulation of pluripotency-associated gene expression. *In vitro*, supplementation with cell-permeable α KG directly supports ES-cell self-renewal while cell-permeable succinate promotes differentiation. This work reveals that intracellular α KG/succinate levels can contribute to the maintenance of cellular identity and have a mechanistic role in the transcriptional and epigenetic state of stem cells.

Mouse ES cells can be maintained in two medium formulations: a serum-free medium reported to support a cellular phenotype that mimics 'naive' epiblast cells of the inner cell mass (containing GSK-3 β and MAPK/ERK inhibitors (2i)/leukaemia inhibitory factor (LIF), hereafter 2i/L); or a serum-based medium that supports the proliferation of a more committed ES cell phenotype (serum/LIF, hereafter S/L)^{4–11}. To characterize ES cell metabolism, we investigated whether cells cultured in these two media have different requirements for glucose and/or glutamine. ES cells cultured in either medium proliferated at equivalent rates when glucose and glutamine were abundant and cells cultured with or without 2i were unable to proliferate in the absence of glucose (Extended Data Fig. 1a, b). In contrast, cells cultured in 2i/L, but not S/L, proliferated robustly in the absence of exogenous glutamine (Fig. 1a and Extended Data Fig. 1c). Likewise, four newly derived ES-cell lines (ESC-1–4) exhibited convincing glutamine-independent proliferation in 2i/L medium while retaining features of pluripotent cells, including ES-cell-like morphology, reactivity to alkaline phosphatase (AP) and the ability to form teratomas (Fig. 1b, c and Extended Data Fig. 1d). Cells cultured in 2i medium alone could also proliferate in the absence of exogenous glutamine (Extended Data Fig. 1e).

This effect was not due to differences in medium nutrient formulations as supplementing S/L medium with the GSK-3 β and MAPK/ERK inhibitors present in 2i also enabled glutamine-independent proliferation while maintaining ES cell morphology and markers of pluripotency (Fig. 1d, e). An alternative ES-cell medium containing BMP4 and LIF added to the same serum-free formulation as in 2i/L¹² failed to support glutamine-independent growth (Fig. 1f). Likewise, epiblast stem cells

(EpiSCs) could not proliferate in the absence of exogenous glutamine (Extended Data Fig. 1f, g). However, the ability to undertake glutamine-independent growth was not limited to embryonic pluripotency; fibroblast-derived induced pluripotent cells (iPSCs) were also able to proliferate in glutamine-free 2i/L medium (Extended Data Fig. 1h). These results indicate that the GSK-3 β and MAPK/ERK inhibitors in 2i-containing medium are both necessary and sufficient to enable proliferation of pluripotent cells in the absence of exogenous glutamine.

The fact that cells proliferated in the absence of exogenous glutamine in 2i/L medium, albeit at a slower rate than cells cultured in glutamine-replete medium (Extended Data Fig. 1i), indicates that these cells must be capable of *de novo* glutamine synthesis. Indeed, chemical inhibition of glutamine synthase was sufficient to block proliferation of cells in glutamine-free 2i/L medium (Extended Data Fig. 1j). Likewise, addition of cell-permeable dimethyl- α -ketoglutarate (DM- α KG), a precursor for glutamine synthesis, was sufficient to enable glutamine-independent proliferation in both S/L and 2i/L conditions (Extended Data Fig. 1k), suggesting that the supply of precursors for glutamine synthesis determines the ability of ES cells to proliferate in the absence of glutamine. In support of this model, cells cultured in 2i/L preserved larger intracellular pools of glutamate after glutamine withdrawal than cells cultured in S/L (Fig. 1g). These results suggest that 2i/L cells can generate

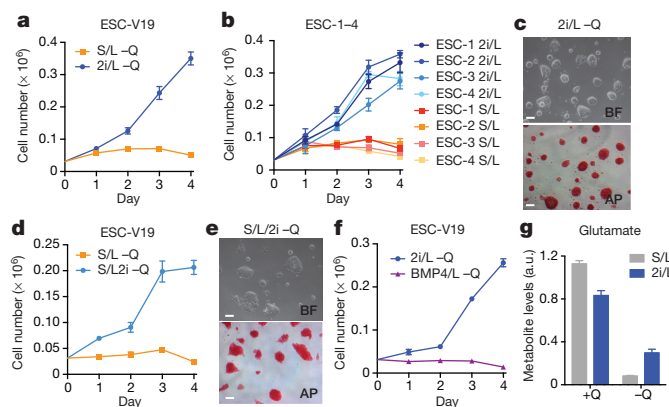


Figure 1 | 2i is necessary and sufficient to confer glutamine independence. **a–f**, Growth curves and representative images of ES cells grown in the absence of glutamine (Q). **a**, **b**, Growth curves of ESC-V19 cells (**a**) and V6.5 ES-cell lines (ESC-1–4) (**b**) cultured in glutamine-free S/L or 2i/L medium. **c**, **e**, Phase images showing ESC-1 cells cultured in glutamine-free 2i/L (**c**) or S/L/2i (**e**) medium for 3 days. Top, brightfield (BF); bottom, AP staining. Scale bars, 500 μ m. **d**, Growth curve of ESC-V19 cells in glutamine-free S/L or S/L/2i medium. **f**, Growth curve of ESC-V19 cells cultured without glutamine in two serum-free medium formulations containing N2 and B27 supplements, 2i/L and BMP4/L. **g**, Intracellular glutamate levels 8 h after addition of medium with or without glutamine. a.u., arbitrary units. Data are presented as the mean \pm standard deviation (s.d.) of triplicate wells from a representative experiment.

¹Laboratory of Chromatin Biology and Epigenetics, The Rockefeller University, New York, New York 10065, USA. ²Cancer Biology and Genetics Program, Memorial Sloan Kettering Cancer Center, New York, New York 10065, USA. ³Donald B. and Catherine C. Marron Cancer Metabolism Center, Memorial Sloan Kettering Cancer Center, New York, New York 10065, USA.

*These authors contributed equally to this work.

glutamate (and glutamine) from carbon sources other than glutamine itself.

Despite their different growth requirements, cells cultured in both S/L and 2i/L consumed high levels of glucose and glutamine and excreted similar levels of lactate, consistent with the metabolic profile of most proliferating cells, including cancer cells and pluripotent cells (Fig. 2a)^{1,13}. Oxidation of glucose and glutamine via the mitochondrial tricarboxylic acid (TCA) cycle provides a critical source of the biosynthetic precursors required for cell proliferation. With the exception of α KG, steady-state levels of TCA cycle metabolites were reproducibly diminished in ES cells cultured in 2i/L (Fig. 2b).

In most cells, glutamine is catabolized to α KG to support TCA cycle anaplerosis (Fig. 2c). ES cells grown in S/L medium exhibited high levels of TCA cycle intermediates and virtually all intracellular glutamate, α KG and malate were rapidly labelled after addition of [U-¹³C]glutamine (Fig. 2d). In contrast, a substantial fraction of these metabolites failed to become labelled with glutamine in ES cells grown in 2i/L. Instead, there was a rapid labelling of these three metabolite pools from [U-¹³C]glucose (Fig. 2e). Quantification of metabolite fluxes revealed that although the flux of glutamine-derived carbons through α KG was similar in both conditions, glutamine flux through malate was significantly diminished in cells cultured in 2i/L, indicating that the entry of glutamine-derived α KG into the TCA cycle is repressed by culture in 2i/L (Fig. 2f). Instead, when cells are cultured in 2i/L, a substantial amount of both α KG and malate was produced from glucose (Fig. 2g).

Consistent with these results, cells cultured with 2i inhibitors demonstrated substantial glucose-dependent glutamate production (Extended Data Fig. 2a). Consequently, during conditions of glutamine depletion,

cells cultured in 2i/L medium were able to use glucose-derived carbons to maintain elevated glutamate pools sufficient to support cell growth (Extended Data Fig. 2b). Moreover, in comparison with their S/L counterparts, 2i/L cells used more glucose-derived carbon and relatively less glutamine-derived carbon to support protein synthesis (Extended Data Fig. 2c), confirming that 2i promotes increased glucose-dependent amino acid synthesis.

Diminished glutamine entry into the TCA cycle, coupled with the observed efflux of glucose-derived carbons from the TCA cycle as glutamate, suggested that cells cultured in 2i/L might not be oxidizing all the α KG produced from glutamine in the mitochondria. Indeed, the α KG/succinate ratio was robustly elevated by 2i/L in every ES-cell line tested (Fig. 3a). Cellular α KG/succinate ratios have been implicated in the regulation of the large family of α KG-dependent dioxygenases¹⁴. As Jumonji C (JmjC)-domain-containing histone demethylases and the Tet family of DNA demethylases comprise a major subset of these enzymes, the elevated ratio of α KG/succinate observed in cells grown in 2i/L medium could have important implications for the regulation of chromatin structure.

Since α KG was largely derived from glutamine metabolism (Fig. 2d), we tested whether glutamine deprivation affected histone lysine methylations known to be regulated in part by α KG-dependent demethylases¹⁵. Cells cultured in glutamine-free medium exhibited increases in trimethylation and decreases in monomethylation on H3K9, H3K27, H3K36 and H4K20, whereas H3K4 methylations remained unchanged (Fig. 3b). DM- α KG reversed the increase in H3K27me3 and H4K20me3 observed in glutamine-deficient medium (Extended Data Fig. 3a), confirming that these changes could be accounted for by the decline in glutamine-dependent

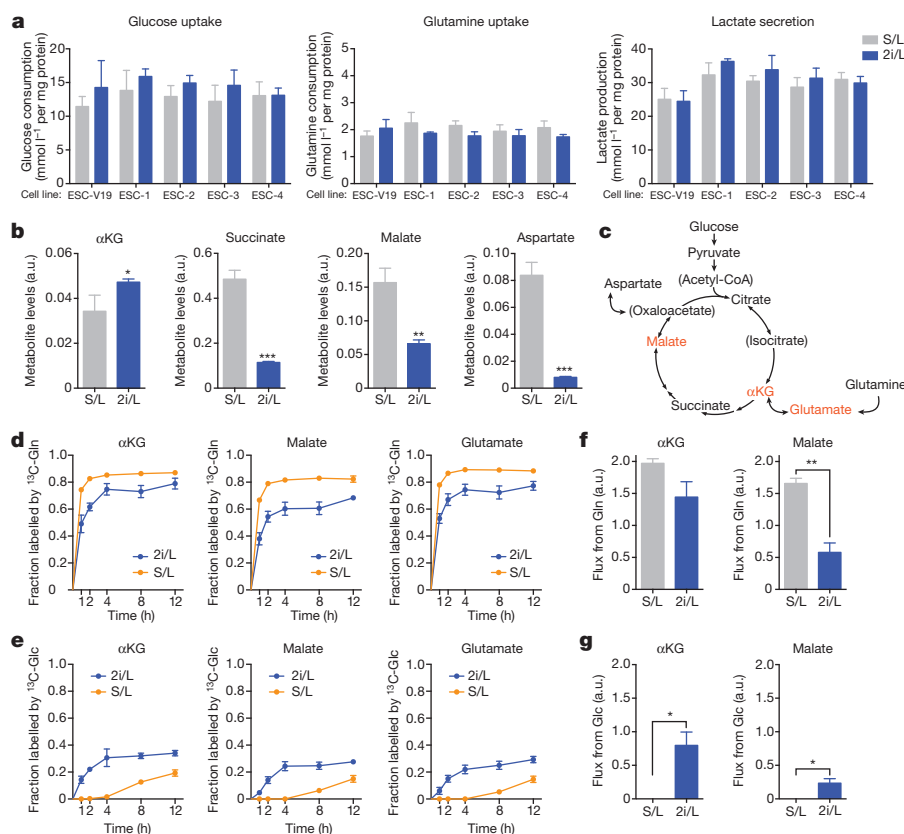


Figure 2 | 2i/L alters glucose and glutamine utilization. **a**, Analysis of glucose uptake (left), glutamine uptake (centre) and lactate secretion (right). **b**, Intracellular metabolite levels. Bars show mean of $n = 4$ (**a**) or $n = 3$ (**b**) replicate wells \pm s.d. from representative experiments. **c**, Schematic of the TCA cycle including entry points for glucose- and glutamine-derived carbons. Isotope tracing was performed for metabolites shown in red. **d**, **e**, Fraction of each metabolite labelled by ¹³C derived from [U-¹³C]glutamine (¹³C-Gln) (**d**)

or derived from [U-¹³C]glucose (¹³C-Glc) (**e**) over time (0–12 h). Mean \pm standard error of the mean (s.e.m.) of three independent experiments are shown. **f**, **g**, Glutamine (**f**) and glucose (**g**) flux through α KG and malate pools. Mean \pm s.e.m. of flux calculated for three independent experiments (shown in **d**, **e**) are shown. * $P < 0.05$, ** $P < 0.005$, *** $P < 0.0005$. P values were determined by unpaired two-tailed Student's t -tests.

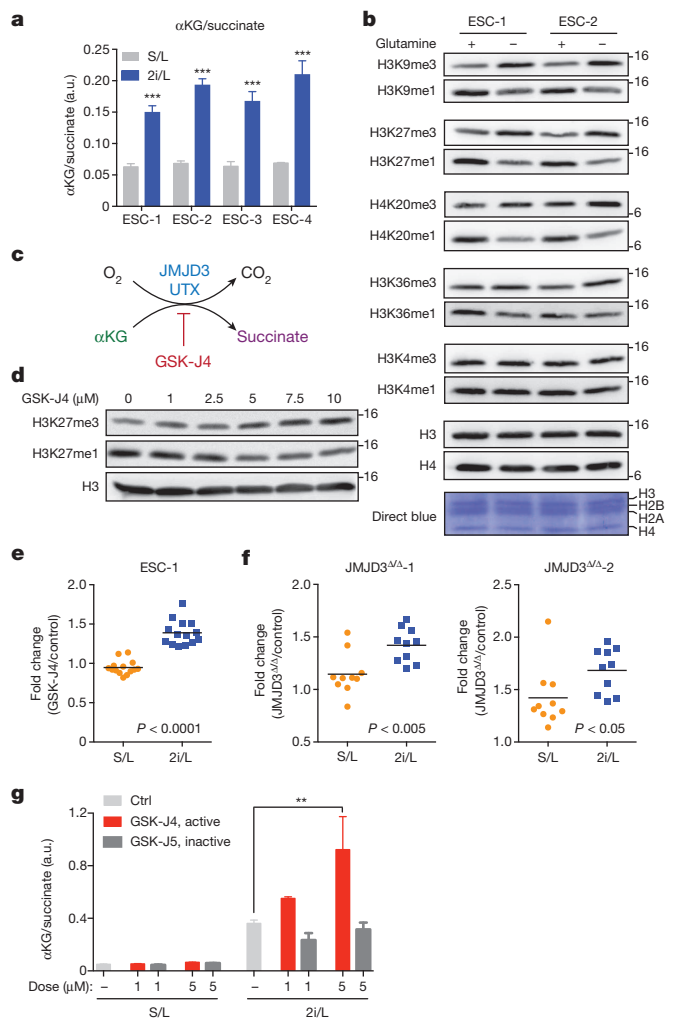


Figure 3 | Histone demethylation is regulated by intracellular α KG in ES cells. **a**, Gas chromatography-mass spectrometry (GC-MS) analysis of the α KG/succinate ratio in ESC-1–4 cells grown in either S/L or 2i/L medium. **b**, Western blot of ESC-1 and ESC-2 cells grown in 2i/L medium with or without glutamine for 3 days. Molecular weight marker (in kDa) is shown. **c**, Simplified schematic of the reaction mechanism of α KG-dependent dioxygenases. **d**, ESC-1 cells grown in S/L in the presence of increasing amounts of GSK-J4 for 24 h. **e**, H3K27me3 chromatin immunoprecipitation followed by quantitative polymerase chain reaction (ChIP-qPCR) of ESC-1 cells cultured in S/L or 2i/L medium containing 30 μ M of GSK-J4 for 5 h. Values represent fold change (GSK-J4/control) at individual bivalent domain genes ($n = 14$). **f**, H3K27me3 ChIP-qPCR of CRISPR/Cas9 edited cells JMJD3 Δ/Δ -1 (left) and JMJD3 Δ/Δ -2 (right) cultured in S/L or 2i/L. Values represent fold change (JMJD3 Δ/Δ cells relative to control cells) at individual bivalent domain genes ($n = 10$). Bars represent mean values. P values were determined by unpaired two-tailed Student's t -test (**e**, **f**). **g**, The ratio of α KG/succinate in ESC-1 cells grown in S/L or 2i/L medium with 1 μ M or 5 μ M of GSK-J4 or GSK-J5 for 3 h. Ctrl, control. ** $P < 0.001$, *** $P < 0.0001$, as determined by two-way analysis of variance (ANOVA) with Sidak's multiple comparisons post-test (**a**, **g**). Data are presented as the mean \pm s.d. (**a**) or s.e.m. (**g**) of triplicate wells from a representative experiment.

α KG. Treatment with GSK-J4 (ref. 16), a cell-permeable inhibitor that preferentially inhibits UTX and JMJD3, the two H3K27me3-specific JmjC-family histone demethylases (Fig. 3c), induced a dose-dependent increase in H3K27me3 with a concomitant reduction of H3K27me1 that was comparable in magnitude to the difference observed when cells were cultured in the presence or absence of glutamine (Fig. 3b, d). These data indicate that the methylation of certain histone lysines, including H3K27, are actively suppressed by α KG-dependent histone demethylases in ES cells maintained in 2i/L medium.

In ES cells, 'bivalent domains' are developmentally regulated genomic regions characterized by the colocalization of H3K4me3 and H3K27me3 (refs 17–19). Recent genome-wide analysis of H3K27me3 in S/L- and 2i/L-cultured ES cells revealed that H3K27me3 was specifically depleted at bivalent domain gene promoters in 2i/L-cultured cells¹¹. Our data suggest that the observed increase in α KG might promote α KG-dependent H3K27me3 demethylation in 2i/L ES cells. Indeed, cells cultured in 2i/L exhibited a greater increase in H3K27me3 at bivalent domain promoters when incubated with the H3K27me3 demethylase inhibitor GSK-J4 than cells cultured in S/L (Fig. 3e and Extended Data Fig. 3b, c). The average fold change across the 14 bivalent promoters tested showed a highly significant increase in 2i/L-cultured ES cells compared with S/L-cultured ES cells (Fig. 3e). Similarly, two independent cell lines with mutations in the Jumonji domain of the H3K27me3 demethylase JMJD3 (JMJD3 Δ/Δ -1 and JMJD3 Δ/Δ -2) (Extended Data Fig. 4a–c) demonstrated increases in H3K27me3 levels relative to control lines that were significantly elevated in cells cultured in 2i/L, reflecting enhanced demethylation at these loci in ES cells cultured in 2i/L (Fig. 3f). Furthermore, treatment with GSK-J4, but not the inactive isomer GSK-J5, increased the α KG/succinate ratio in cells cultured in 2i/L (Fig. 3g). These results indicate that 2i/L rewires glutamine metabolism to maintain α KG pools favouring active demethylation of a variety of histone marks.

In addition to reduced H3K27me3 at bivalent domain promoters, cells cultured in 2i/L exhibit DNA hypomethylation^{5,7–9}. Incubating cells with ascorbic acid, a cofactor for α KG-dependent dioxygenases, activates Tet-dependent gene expression and promotes DNA demethylation²⁰. Therefore, we tested whether α KG treatment could exert similar effects (Extended Data Fig. 5a). Total DNA methylation was reduced in cells cultured with cell-permeable α KG (Extended Data Fig. 5b) and treatment with α KG, but not succinate, induced expression of inner-cell-mass- and germline-associated genes previously identified as targets of Tet-mediated activation (Extended Data Fig. 5c)^{20,21}. The effects of α KG persisted upon extended passaging (Extended Data Fig. 5d) and were largely abrogated in *Tet1/Tet2* double-knockout ES cells (Extended Data Fig. 5e). These results suggest that intracellular α KG production may stimulate the activity of multiple α KG-dependent dioxygenases to regulate coordinately the epigenetic marks characteristic of naive pluripotency.

To test whether modulation of the α KG/succinate ratio can influence pluripotent cell fate decisions, we performed colony-formation assays with S/L-cultured ES cells in the presence of α KG or succinate. Colonies formed in S/L medium supplemented with DM- α KG (S/L+DM- α KG) had brighter AP staining and retained the compact colony morphology typical of undifferentiated ES cells (Fig. 4a). Although the total number of colonies was similar in all three conditions, the S/L+DM- α KG wells contained more than double the number of fully undifferentiated colonies compared with S/L only and S/L+DM-succinate (Fig. 4b). As a further test of the ability of α KG to promote maintenance of ES cells, we used a knock-in Nanog-green fluorescent protein (GFP) reporter line²² and found that α KG was sufficient to enhance Nanog expression in a dose-dependent manner (Fig. 4c and Extended Data Fig. 6). These results support the conclusion that α KG promotes the self-renewal of ES cells *in vitro*.

These data demonstrate that the cellular α KG/succinate ratio contributes to the ability of ES cells to suppress differentiation. The rewiring of cellular metabolism by inhibitors of GSK-3 β and MAPK/ERK signalling results in a reprogramming of glucose and glutamine metabolism; in turn, this leads to accumulation of α KG and favours demethylation of repressive chromatin marks such as DNA methylation and H3K9me3, H3K27me3 and H4K20me3 (see Supplementary Discussion). Future studies will investigate the mechanisms through which these inhibitors influence the nuclear/cytosolic accumulation of α KG derived from glucose and glutamine. While we cannot rule out chromatin-independent effects of α KG supplementation on ES cells, our results support the notion that chromatin in pluripotent ES cells is responsive to alterations in intracellular metabolism. Indeed, recent clonal analysis

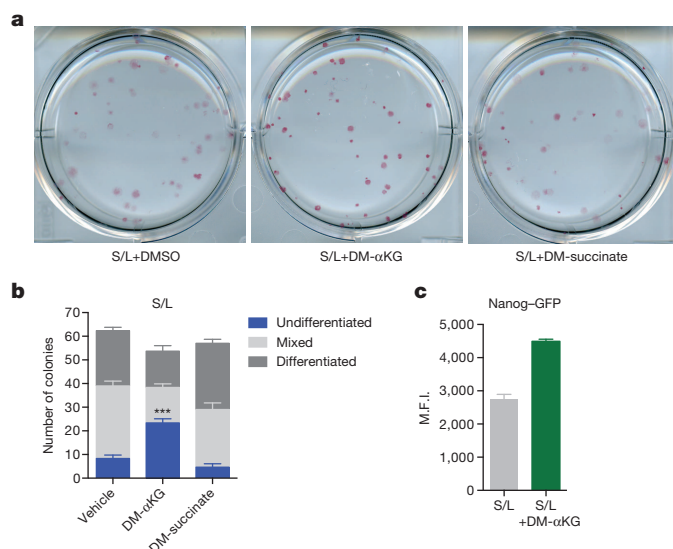


Figure 4 | α KG promotes the maintenance of pluripotency. **a, b,** Colony formation assay using ESC-1 cells. Cells were plated at clonal density and media were changed to experimental media containing either DM- α KG or DM-succinate on day 2 and then analysed 4 days later by AP staining and scored for number of differentiated, mixed and undifferentiated colonies.

a, Representative brightfield images of AP-stained colonies. Vehicle, DMSO, dimethylsulphoxide. **b,** Quantification of colonies. DM- α KG has more undifferentiated colonies than vehicle- or DM-succinate-treated wells. *** $P < 0.0001$, calculated by two-way ANOVA with Tukey's multiple comparisons post-test. **c,** Mean GFP intensity of Nanog-GFP cells treated for 3 days with or without DM- α KG. M.F.I., mean fluorescence intensity. Data are presented as the mean \pm s.e.m. (**b**) or 95% confidence intervals (**c**) of triplicate wells from a representative experiment.

of pluripotent cells revealed that DNA methylation is highly dynamic, balancing the antagonistic processes of removal and addition²³. Together, these results suggest that continued elucidation of the interconnections between signal transduction and cellular metabolism will shed important light on stem cell biology, organismal development and cellular differentiation.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 29 April; accepted 20 October 2014.

Published online 10 December 2014.

1. Lunt, S. Y. & Vander Heiden, M. G. Aerobic glycolysis: meeting the metabolic requirements of cell proliferation. *Annu. Rev. Cell Dev. Biol.* **27**, 441–464 (2011).
2. Eagle, H., Oyama, V. I., Levy, M., Horton, C. L. & Fleischman, R. The growth response of mammalian cells in tissue culture to L-glutamine and L-glutamic acid. *J. Biol. Chem.* **218**, 607–616 (1956).
3. Ying, Q. L. *et al.* The ground state of embryonic stem cell self-renewal. *Nature* **453**, 519–523 (2008).

4. Nichols, J., Silva, J., Roode, M. & Smith, A. Suppression of Erk signalling promotes ground state pluripotency in the mouse embryo. *Development* **136**, 3215–3222 (2009).
5. Smith, Z. D. *et al.* A unique regulatory phase of DNA methylation in the early mammalian embryo. *Nature* **484**, 339–344 (2012).
6. Wray, J., Kalkan, T. & Smith, A. G. The ground state of pluripotency. *Biochem. Soc. Trans.* **38**, 1027–1032 (2010).
7. Leitch, H. G. *et al.* Naive pluripotency is associated with global DNA hypomethylation. *Nature Struct. Mol. Biol.* **20**, 311–316 (2013).
8. Ficiz, G. *et al.* FGF signaling inhibition in ESCs drives rapid genome-wide demethylation to the epigenetic ground state of pluripotency. *Cell Stem Cell* **13**, 351–359 (2013).
9. Habibi, E. *et al.* Whole-genome bisulfite sequencing of two distinct interconvertible DNA methylomes of mouse embryonic stem cells. *Cell Stem Cell* **13**, 360–369 (2013).
10. Borgel, J. *et al.* Targets and dynamics of promoter DNA methylation during early mouse development. *Nature Genet.* **42**, 1093–1100 (2010).
11. Marks, H. *et al.* The transcriptional and epigenomic foundations of ground state pluripotency. *Cell* **149**, 590–604 (2012).
12. Ying, Q. L., Nichols, J., Chambers, I. & Smith, A. BMP induction of Id proteins suppresses differentiation and sustains embryonic stem cell self-renewal in collaboration with STAT3. *Cell* **115**, 281–292 (2003).
13. Zhang, J., Nuebel, E., Daley, G. Q., Koehler, C. M. & Teitell, M. A. Metabolic regulation in pluripotent stem cells during reprogramming and self-renewal. *Cell Stem Cell* **11**, 589–595 (2012).
14. Kaelin, W. G. Jr. Cancer and altered metabolism: potential importance of hypoxia-inducible factor and 2-oxoglutarate-dependent dioxygenases. *Cold Spring Harb. Symp. Quant. Biol.* **76**, 335–345 (2011).
15. Cloos, P. A., Christensen, J., Agger, K. & Helin, K. Erasing the methyl mark: histone demethylases at the center of cellular differentiation and disease. *Genes Dev.* **22**, 1115–1140 (2008).
16. Kruidenier, L. *et al.* A selective jumoni H3K27 demethylase inhibitor modulates the proinflammatory macrophage response. *Nature* **488**, 404–408 (2012).
17. Bernstein, B. E. *et al.* A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**, 315–326 (2006).
18. Boyer, L. A. *et al.* Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature* **441**, 349–353 (2006).
19. Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553–560 (2007).
20. Blaschke, K. *et al.* Vitamin C induces Tet-dependent DNA demethylation and a blastocyst-like state in ES cells. *Nature* **500**, 222–226 (2013).
21. Hackett, J. A. *et al.* Synergistic mechanisms of DNA demethylation during transition to ground-state pluripotency. *Stem Cell Reports* **1**, 518–531 (2013).
22. Faddah, D. A. *et al.* Single-cell analysis reveals that expression of Nanog is biallelic and equally variable as that of other pluripotency factors in mouse ESCs. *Cell Stem Cell* **13**, 23–29 (2013).
23. Shipony, Z. *et al.* Dynamic and static maintenance of epigenetic memory in pluripotent and somatic cells. *Nature* **513**, 115–119 (2014).

Supplementary Information is available in the online version of the paper.

Acknowledgements B.W.C. is a Howard Hughes Medical Institute fellow of the Jane Coffin Childs Memorial Research Fund. L.W.S.F. is the Jack Sorrell Fellow of the Damon Runyon Cancer Research Foundation (DRG-2144-13). This work was funded by grants from the National Institutes of Health/National Institute of General Medical Sciences (C.D.A.) and from the National Cancer Institute (C.B.T.). We thank C. Li for assistance with FACS analysis and M. Dawlaty, D. Faddah and R. Jaenisch for sharing cell lines used in this study.

Author Contributions B.W.C. and L.W.S.F. designed and performed all experiments in the study under the guidance of C.D.A. and C.B.T. J.R.C. contributed material support. B.W.C., L.W.S.F., C.D.A. and C.B.T. wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to C.B.T. (thompsonc@mskcc.org) or C.D.A. (alliscd@mail.rockefeller.edu).

METHODS

Cell lines. ESC-1–4 lines are V6.5 ES cells derived from C57BL/6 \times 129S4/SvJae F1 embryos in 2i/L medium. Cells were derived from embryonic day (E)3.5 blastocysts following standard ES-cell isolation procedures²⁴. Flushed blastocysts were plated onto laminin-coated dishes (20 μ g ml⁻¹, Stemgent 06-0002) in 2i/L medium. Mice were purchased from Jackson Laboratories (C57BL/6 JAX 000664 and 129S4/SvJae JAX 009104). *Tet1/2* double-knockout ES cells²⁵, V19 ES cells (ESC-V19) and OKS iPSCs²⁶ were a gift from R. Jaenisch. All cells were routinely tested for mycoplasma contamination. Mice were maintained at The Rockefeller University. All animal procedures were designed following National Institutes of Health guidelines and approved by the Institutional Animal Care and Use Committee at The Rockefeller University.

Cell culture. Maintenance media for ES cells were as follows: serum/LIF (S/L) maintenance medium contained Knockout DMEM (Gibco) supplemented with 15% ES-cell-qualified FBS (Gemini), penicillin/streptomycin (Life Technologies), 0.1 mM 2-mercaptoethanol, 2 mM L-glutamine (Life Technologies) and LIF plated onto irradiated feeder mouse embryonic fibroblasts (MEFs); 2i/LIF (2i/L) maintenance conditions used a base medium made from a 1:1 mix of DMEM/F12 (Life Technologies 11302-033) and Neurobasal medium (Life Technologies 21103-049) containing N2 and B27 supplements (Life Technologies 17502-048 and 17504-044, 1:100 dilutions), penicillin/streptomycin, 0.1 mM 2-mercaptoethanol, 2 mM L-glutamine, LIF, CHIR99021 at 3 μ M (Stemgent) and PD0325901 at 1 μ M (Stemgent). Experimental media used for all experiments (except growth curves without glucose, ¹³C isotope tracing experiments and ¹⁴C labelling experiments) contained a 1:1 mix of glutamine-free DMEM (Life Technologies 11960-051) and Neurobasal medium (Life Technologies 21103-049) with or without 2 mM glutamine. With the exception of 15% dialysed FBS (Gemini 100-108) in S/L experimental medium, all other supplements were equivalent to maintenance media (S/L or 2i/L). For growth curves without glucose, ¹³C isotope tracing experiments and ¹⁴C labelling experiments, media contained a 1:1 mix of glutamine- and glucose-free DMEM (Invitrogen A14430-01) and glutamine- and glucose-free Neurobasal medium (Invitrogen 0050128DJ) containing either 20 mM [U-¹³C]glucose or 2 mM [U-¹³C]glutamine (Cambridge Isotope Laboratories) and either 20 mM unlabelled glucose or 2 mM unlabelled glutamine as necessary; all supplements were the same as in experimental media described above (S/L or 2i/L). All experiments were performed using feeder-free conditions. ESC-1 EpiSCs were cultured feeder-free on fibronectin (Sigma)-coated plates in EpiSC maintenance medium including DMEM/F12, N2 and B27 supplements, penicillin/streptomycin, 0.1 mM 2-mercaptoethanol, L-glutamine, 75 μ g ml⁻¹ BSA (Gibco) supplemented with human activin A (20 ng ml⁻¹; Peprotech) and bFGF (10 ng ml⁻¹; Invitrogen). EpiSCs were passaged 1:2 or 1:4 using Accutase every other day. For ES-cell to EpiSC differentiation, ESC-1 cells were plated onto fibronectin-coated dishes. Twenty-four hours after plating, the medium was changed to EpiSC maintenance medium supplemented with 6 μ M JAK inhibitor (Calbiochem) for five passages. Analysis was performed on passage 7 EpiSCs. GSK-J4 and GSK-J5 were purchased from Tocris Bioscience.

Teratomas. ESC-1 cells were plated in maintenance medium at a concentration of 2.5×10^5 cells per T25 dish. The following day medium was changed to 2i/L experimental medium with or without glutamine. 72 h later, 1×10^6 cells were harvested from each group and mixed 1:1 with experimental medium (without glutamine) plus Matrigel Basement Membrane Matrix (BD) or experimental medium alone and injected into the flanks of recipient SCID mice aged 8–12 weeks (NOD scid gamma JAX 005557 purchased from Jackson Laboratories). All conditions produced tumours in 4–8 weeks. Mice were euthanized before tumour size exceeded 1.5 cm in diameter. Tumours were excised and fixed in 4% paraformaldehyde overnight at 4 °C. Tumours were paraffin-embedded and sections were stained with haematoxylin and eosin according to standard procedures by Histoserv.

Glucose, glutamine and lactate measurements. Glucose, glutamine and lactate levels in culture medium were measured using a YSI 7100 multichannel biochemistry analyser (YSI Life Sciences). Fresh medium was added to 12-well plates of sub-confluent cells and harvested 48 h later. Changes in metabolite concentrations relative to fresh media were normalized to the protein content of each well. These experiments were performed independently at least two times.

Metabolite profiling. For all metabolite experiments, cells were seeded in their standard culture medium in 6-well plates and the next day were changed into experimental medium. Medium was changed again at the indicated time before harvest (usually 1–24 h). Metabolites were extracted with 1 ml ice-cold 80% methanol supplemented with 20 μ M deuterated 2-hydroxyglutarate (D-2-hydroxyglutaric-2,3,3,4,4-d₅ acid (d5-2HG)) as an internal standard. After overnight incubation at -80 °C, lysates were harvested and centrifuged at 21,000g for 20 min to remove protein. Extracts were dried in an evaporator (Genevac EZ-2 Elite) and resuspended by incubation at 30 °C for 2 h in 50 μ l of 40 mg ml⁻¹ methoxyamine hydrochloride in pyridine. Metabolites were further derivatized by addition of 80 μ l of MSTFA plus 1% TMCS (Thermo Scientific) and 70 μ l ethyl acetate (Sigma) and incubated at

37 °C for 30 min. Samples were analysed using an Agilent 7890A GC coupled to Agilent 5975C mass selective detector. The GC was operated in splitless mode with constant helium gas flow at 1 ml min⁻¹. One microlitre of derivatized metabolites was injected onto an HP-5MS column and the GC oven temperature ramped from 60 °C to 290 °C over 25 min. Peaks representing compounds of interest were extracted and integrated using MassHunter software (Agilent Technologies) and then normalized to both the internal standard (d5-2HG) peak area and the protein content of duplicate samples as determined by a BCA protein assay (Thermo Scientific). Ions used for quantification of metabolite levels are as follows: d5-2HG *m/z* 354; α KG, *m/z* 304; aspartate, *m/z* 334; glutamate, *m/z* 363; malate, *m/z* 335; and succinate, *m/z* 247. All peaks were manually inspected and verified relative to known spectra for each metabolite. For isotope tracing studies, experiments were set up as described earlier using glucose- and glutamine-free DMEM: NB media base supplemented with ¹²C-glucose (Sigma) and ¹²C-glutamine (Gibco) or the ¹³C versions of each metabolite, [U-¹³C]glucose or [U-¹³C]glutamine (Cambridge Isotope Laboratories). Enrichment of ¹³C was assessed by quantifying the abundance of the following ions: α KG, *m/z* 304–315; aspartate, *m/z* 334–346; glutamate, *m/z* 363–377; and malate, *m/z* 335–347. Correction for natural isotope abundance was performed using IsoCor software²⁷. Flux was calculated as the product of the first order rate constant of the kinetic labelling curve and relative metabolite pool size (normalized to mean S/L values for each experiment)²⁸. The flux from glucose- and glutamine-derived carbons was calculated for each of three independent experiments and the average flux for each metabolite was shown. Flux experiments represent the average of three independent experiments; all other experiments were performed independently at least twice and a representative experiment is shown.

Protein labelling. ES cells were plated at 7.5×10^5 cells per 6-well plate into experimental medium (S/L or 2i/L) containing 0.01% unenriched D-[U-¹⁴C]-glucose (Perkin Elmer NEC042V250UC) or L-[U-¹⁴C]-glutamine (Perkin Elmer NEC45 1050UC). Forty-eight hours later, cells were washed with PBS, scraped and pelleted at 4 °C. Protein pellets devoid of lipid fractions were isolated according to the Bligh-Dyer method²⁹. Briefly, pellets were resuspended in 200 μ l distilled H₂O, 265 μ l 100% methanol and 730 μ l of chloroform. Samples were vortexed for 1 h at 4 °C. The organic phase was removed and the remaining sample washed with 1 \times volume of methanol and spun at 14,200g for 5 min. The supernatant was discarded and the pellet was resuspended in 6 M guanidine hydrochloride at 65 °C for 30–45 min. Samples were quantified using Beckman LS 6000IC instrument. Values represent the average from four wells normalized to protein of duplicate samples. Labelling experiments were performed twice.

Growth curves. ES cells or EpiSCs were plated in maintenance medium at a concentration of 375,000 cells per 12-well plate. The following day cells were washed with PBS and media were changed to experimental media (for S/L conditions this included dialysed FBS) with or without individual metabolites. Cells were counted each day using a Beckman Coulter Multisizer 4. All growth curves were performed independently at least two times.

ChIP. Native ChIP assays (histones) were performed with approximately 6×10^6 ES cells per experiment. Cells were subject to hypotonic lysis and treated with micrococcal nuclease to recover mono- to tri-nucleosomes. Nuclei were lysed by brief sonication and dialysed into N-ChIP buffer (10 mM Tris pH 7.6, 1 mM EDTA, 0.1% SDS, 0.1% Na-Deoxycholate, 1% Triton X-100) for 2 h at 4 °C. Soluble material was incubated overnight at 4 °C after addition of 0.5–1 μ g of antibody bound to 25 μ l protein A Dynal magnetic beads (Invitrogen), with 5% kept as input DNA. Magnetic beads were washed, chromatin was eluted and ChIP DNA was dissolved in 10 mM Tris pH 8 for quantitative PCR reactions (see later). Three separate ChIP experiments were performed on replicate biological samples. The data shown are the average qRT-PCR values ($n = 3$).

ChIP-qPCR. Primers are listed below. All qPCR was performed using an Applied Biosystems StepOnePlus system and Power SYBR Green PCR master mix. ChIP samples were diluted 1:100 in H₂O and 5 μ l was used per reaction. ChIP-qPCR signals were calculated as per cent input. Primers were as follows: *Gata6*, forward, 5'-CGCAGCACACAGGTACAGTT-3', reverse, 5'-GGGATCCAAGCAGATTGA AA-3'; *Pax9*, forward, 5'-AGGTGTGCGACAGCTAAAGG-3', reverse, 5'-ATC AACCCGGAGTGATCAAG-3'; *Lhx1*, forward, 5'-TGCCAGGCACCATTA CA GT-3', reverse, 5'-AGGCAAAGGAAACCATGA-3'; *Hoxa2*, forward, 5'-CC AATGACAATTGGGCTTT-3', reverse, 5'-TGAGGCGTTCTTCTGACT-3'; *Hoxc9*, forward, 5'-TTCTTCCCTTTGGCCTTTT-3', reverse, 5'-AGGGTGTC TTGGCTCTCTCA-3'; *Evs1*, forward, 5'-GCCAGGTGATCTGGGTGGGGA-3', reverse, 5'-TGAGAACCCGGCCTTGTGTGCT-3'; *Fgf5*, forward, 5'-GGGATCTC CTGTGCTGGGGT-3', reverse, 5'-AGGCCTGTACTGCAGCCACATT-3'; *Ascl2*, forward, 5'-GCTCCAGAAGCAGTCTCCCTGA-3', reverse, 5'-GATA GAGCCAGAGCCCAAGCCCC-3'; *Lrat*, forward, 5'-CCAAGTCCTTCAGTCT CTTGCCCC-3', reverse, 5'-GGCCACACAGGCTGCTTCCA-3'; *Lhx5*, forward, 5'-AACCTTAGGCCCCAGCCCC-3', reverse, 5'-CGTGGGCTGGAGGGG AGAA-3'; *Sox17*, forward, 5'-GTCTCCCCATGTAGCTCTCTGCC-3', reverse,

5'-AGAAGAGTCACTGTGGAGGTGAGGG-3'; *brachyury*, forward, 5'-GCCA CTGCTTTCCCGAGACCC-3'; reverse, 5'-CCAGGACAGGCAGGTAGGGG-3'; *Gata4*, forward, 5'-ACGTGTGGTGTAAATGTGCAAGCC-3'; reverse, 5'-TGCC CACAAGCCTGCGATCC-3'; *Sox21*, forward, 5'-AACAGACATGCCAGTCAG CAGTG-3'; reverse, 5'-TTAGCATCGCACCACCCAGAGTC-3'; *Pou5f1*, forward, 5'-GAGGTCAAGGCTAGAGGGTGG-3'; reverse 5'-AGGGACGGTTTC ACCTCTCC-3'.

qRT-PCR. RNA was isolated using the RNeasy kit (Qiagen). After DNase treatment, 1–2 µg RNA was used for cDNA synthesis using the First-Strand Synthesis kit (Invitrogen). Quantitative RT-PCR analysis was performed in biological triplicate using an ABI Prism 7000 (Applied Biosystems) with Platinum SYBR green. All data were generated using cDNA from three wells for each condition. Primers used were as follows: *Pou5f1*, forward, 5'-ACATCGCCAATCAGCTTGG-3'; reverse, 5'-AGAACCATACTGAACCACATCC-3'; *Nanog*, forward, 5'-AAGATGCGG ACTGTGTTCTC-3'; reverse, 5'-CGCTGCACTTCATCCTTTG-3'; *Esrrb*, forward, 5'-TTTCTGGAACCCATGGAGAG-3'; reverse, 5'-AGCCAGCACCTCC TTCTACA-3'; *Klf2*, forward, 5'-TAAAGGCGCATCTGCGTACA-3'; reverse, 5'-CGCACAAGTGGCACTGAAAG-3'; *Nr0b1*, forward, 5'-TCCAGGCCATCAA GAGTTTC-3'; reverse, 5'-ATCTGCTGGGTTCTCCACTG-3'; *Fgf5*, forward, 5'-AAACTCCATGCAAGTGCCAAAT-3'; reverse, 5'-TCTCGGCCTGTCTTTTC AGTTC-3'; *Zfp42*, forward, 5'-CGAGTGGCAGTTTCTTCTTGG-3'; reverse, 5'-CTTCTTGAACAATGCCTATGACTCACTTCC-3'; *actin*, forward, 5'-TGGCG CTTTTGACTCAGGAT-3'; reverse, 5'-GGGATGTTTGTCCAAACAA-3'; *Asz1*, forward, 5'-GAGTGGGCTTCTCCAGAAA-3'; reverse, 5'-GGTCATTTTCCC GCTCATTC-3'; *Wdrl5a*, forward, 5'-TGTGTGGAACCTGGACAAC-3'; reverse, 5'-GCCAATGCCGTCGTTATTTT-3'; *Daz1*, forward, 5'-CAACTGTAACTAC CACTGCAG-3'; reverse, 5'-CAAGAGACCACTGTCTGTATGC-3'; *Gapdh*, forward, 5'-TTCACCACCATGGAGAAGGC-3'; reverse, 5'-CCCTTTGGCTCCA CCCT-3'.

DNA methylation. Genomic DNA was extracted from ES-cell samples using Puregene Core Kit A (Sigma). DNA methylation was measured using the colorimetric MethylFlash Methylated DNA quantification kit (Epigentek) according to manufacturer instructions. ELISA experiments were performed independently two times.

CRISPR/Cas9 ES cells. A Cas9-2A-Puro plasmid was purchased from Addgene (Addgene plasmid 48139)³⁰. Two gRNAs targeting exon 17 of mouse *Jmjd3* were designed using the online software (<http://crispr.mit.edu>) resource from the Zhang Laboratory and were cloned into Cas9-2A-Puro using the BbsI restriction enzyme sites. ESC-1 cells cultured in 2i/L medium were transfected with either Cas9-2A-Puro control or *Jmjd3* gRNA-containing plasmids using Lipofectamine 2000 (Life Technologies). After 24 h, cells were changed to fresh medium containing 1 µg ml⁻¹ puromycin for 48 h. After selection, cells were cultured for 24 h in 2i/L medium and then split to clonal density. After approximately 7 days, colonies were picked and expanded for analysis. Genomic DNA was purified from individual clones and used for PCR amplification of regions surrounding each gRNA target site. gRNA #1 product is 367 bp and gRNA #2 is 317 bp. Cloning of PCR products was performed using pGEM-T Easy (Promega). Mutants were identified by Sanger sequencing (Genewiz). gRNA oligonucleotides were as follows: *Jmjd3* gRNA #1, forward, 5'-CACCTGTGGATGTTACCCGCATGA-3'; reverse, 5'-AAACTCATGCGGG TAACATCCACA-3'; *Jmjd3* gRNA #2, forward, 5'-CACCGTCCCTGGCAGCC GAACGCC-3'; reverse, 5'-AAACGGCGTTTCGGCTGCCAGGGAC-3'. PCR primers were as follows: *Jmjd3* gRNA #1, forward, 5'-GGCTAAGGCCTAAGAGT

GCG-3'; reverse, 5'-CGGACCCCAAGAACCATCAC-3'; *Jmjd3* gRNA #2, forward, 5'-TGGCCTGCAGAGGGAGATAG-3'; reverse, 5'-ATTTCTCGGCAT TCCTGTG-3'.

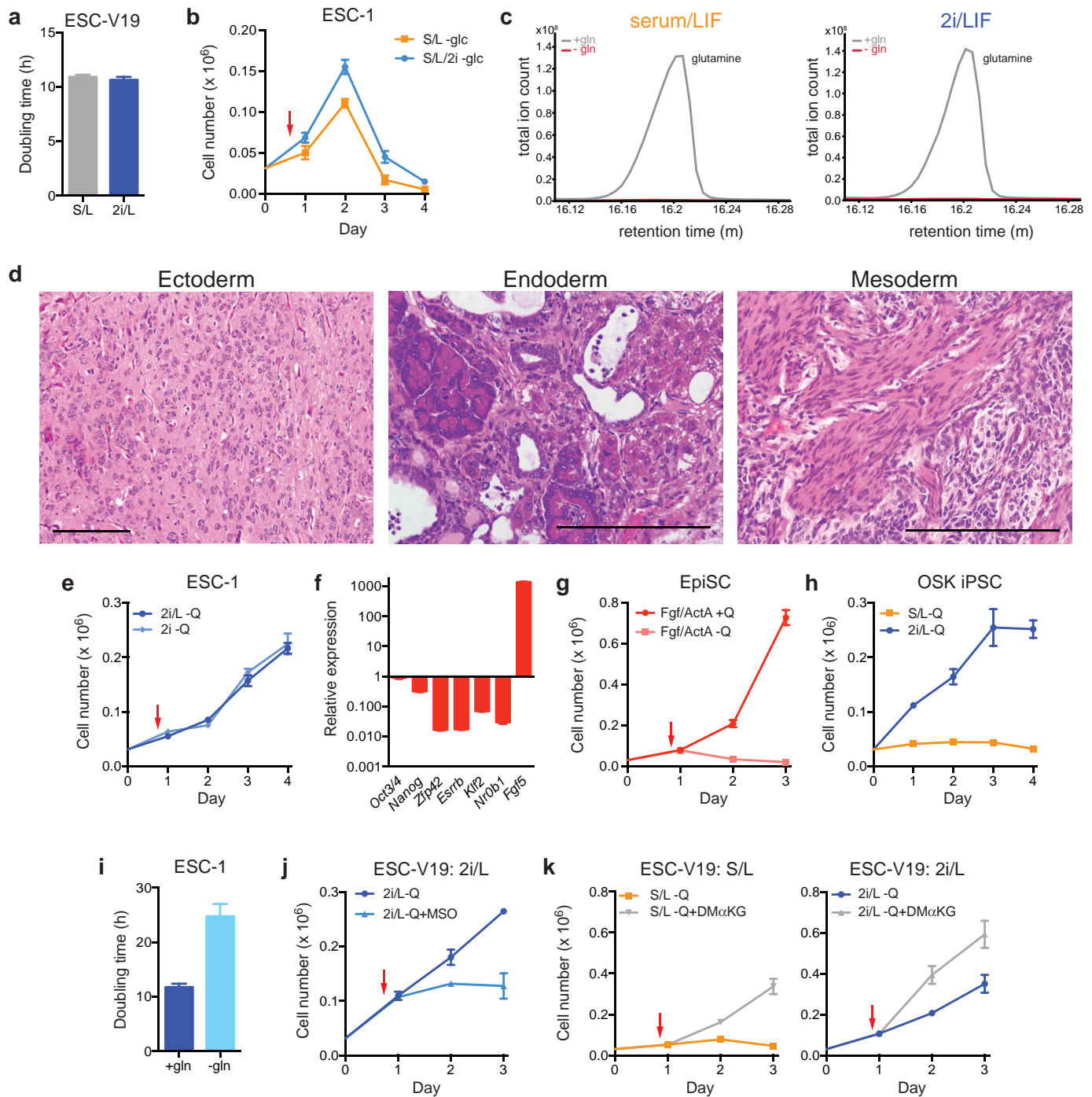
FACS. Nanog-GFP ES cells²² were cultured in S/L experimental medium for three passages and 2.5 × 10⁴ cells were plated into a 6-well plate. Twenty-four hours later medium was changed to S/L medium containing vehicle control or DM-αKG. Media were subsequently changed 48 h later and cells were harvested the following day. FACS analysis was performed at The Rockefeller University Flow Cytometry Resource Center using a BD LSR II. Data were generated using FlowJo. Experiments were performed two independent times and a representative experiment depicting triplicate biological wells is shown.

Western blot analysis. Lysates were extracted in 1× Laemmli buffer, separated by SDS-PAGE and transferred to Immobilon PVDF (Millipore) membranes. Membranes were blocked in 5% milk prepared in phosphate-buffered saline (PBS) plus 0.1% Tween 20 (PBS-T), incubated with primary antibodies overnight at 4 °C and with horseradish peroxidase (HRP)-conjugated secondary antibodies for 1 h the next day. After ECL application (Millipore), imaging was performed using Lumimager LAS-3000 (FujiFilm). The following antibodies were used for western blotting: H3 (Abcam 1791), H3K4me3 (Active Motif 39159), H3K4me1 (Millipore 07-436), H3K9me1 (gift from T. Jenuwein), H3K9me3 (Active Motif 39161), H4 (Abcam 0158), H4K20me1 (Abcam 9051), H4K20me3 (Millipore 07-463), H3K27me1 (Millipore 07-448), H3K27me3 (Millipore 07-449), H3K36me3 (Abcam 9050) and H3K36me1 (Millipore 07-548). All antibodies were used at a dilution of 1:1,000. H3K27me3 antibody used for ChIP-qPCR, Cell Signaling 9733BF.

Self-renewal assays. ES cells free from feeder MEFs were plated at 100 cells per well in 6-well plates coated with 20 µg ml⁻¹ mouse laminin (Stemgent 06-0002) in maintenance S/L medium. The next day, media were changed to S/L experimental medium containing dimethyl-α-ketoglutarate (4 mM, Sigma 349631), dimethylsuccinate (4 mM, Sigma W239607) or DMSO vehicle control. Four days later, cells were washed with PBS and stained for alkaline phosphatase using Vector Red Alkaline Phosphatase Kit (Vector Labs) according to manufacturer's instructions. Self-renewal assays were performed independently at least two times.

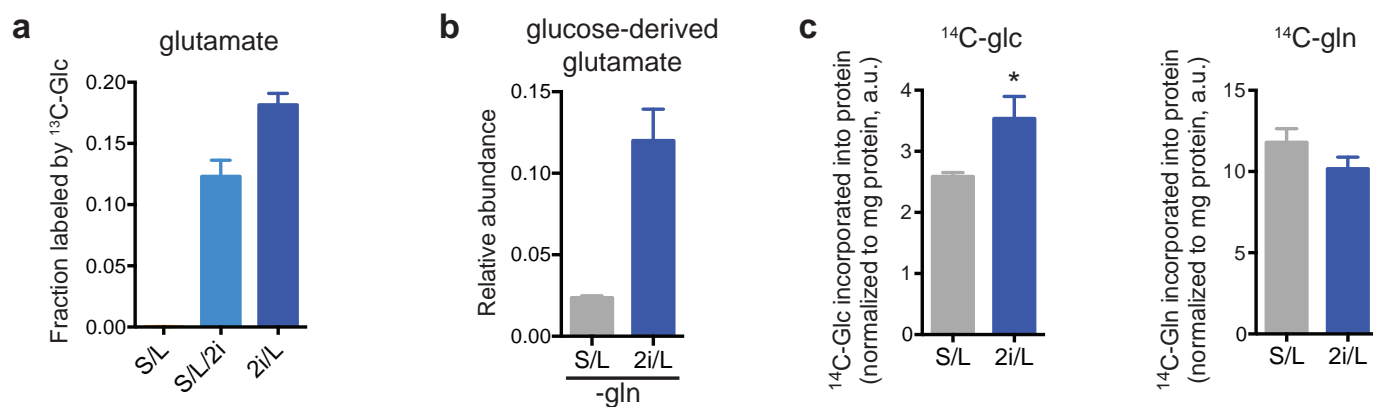
Statistics. Comparisons were made using unpaired two-tailed Student's *t*-tests or two-way ANOVA with appropriate post-test (determined using GraphPad Prism) as indicated. Experiments were performed with three or four replicates as is the standard in the field. Variation is shown as s.d., s.e.m. or 95% confidence intervals as indicated in figure legends.

24. Markoulaki, S., Meissner, A. & Jaenisch, R. Somatic cell nuclear transfer and derivation of embryonic stem cells in the mouse. *Methods* **45**, 101–114 (2008).
25. Dawlaty, M. M. *et al.* Combined deficiency of Tet1 and Tet2 causes epigenetic abnormalities but is compatible with postnatal development. *Dev. Cell* **24**, 310–323 (2013).
26. Buganim, Y. *et al.* Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. *Cell* **150**, 1209–1222 (2012).
27. Millard, P., Letisse, F., Sokol, S. & Portais, J. C. IsoCor: correcting MS data in isotope labeling experiments. *Bioinformatics* **28**, 1294–1296 (2012).
28. Yuan, J., Bennett, B. D. & Rabinowitz, J. D. Kinetic flux profiling for quantitation of cellular metabolic fluxes. *Nature Protocols* **3**, 1328–1340 (2008).
29. Bligh, E. G. & Dyer, W. J. A rapid method of total lipid extraction and purification. *Can. J. Biochem. Physiol.* **37**, 911–917 (1959).
30. Ran, F. A. *et al.* Genome engineering using the CRISPR-Cas9 system. *Nature Protocols* **8**, 2281–2308 (2013).



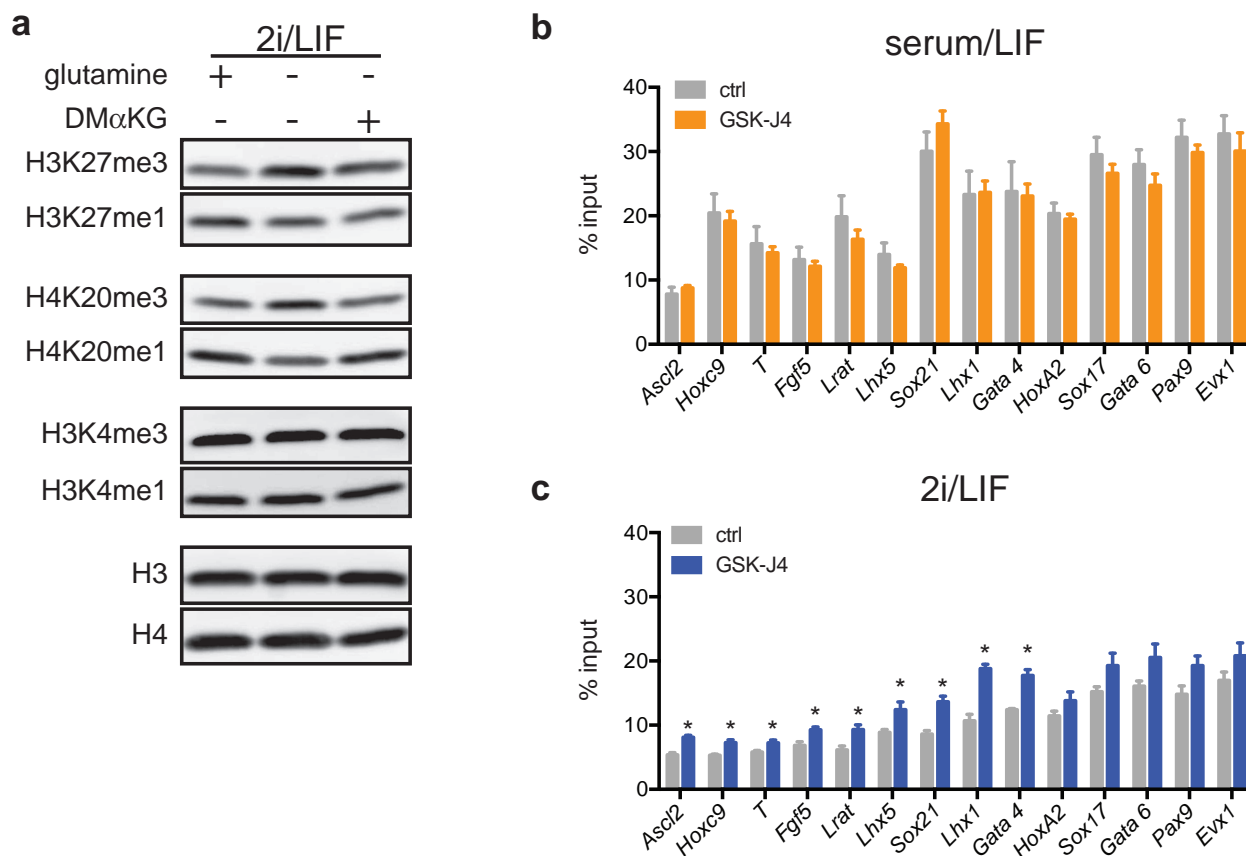
Extended Data Figure 1 | Pluripotent stem cells can proliferate in the absence of glutamine when cultured in 2i/LIF medium. **a**, Doubling time of ESC-V19 cells cultured in S/L or 2i/L. **b**, Growth curve of ESC-1 cells cultured in S/L or S/L/2i medium devoid of glucose. **c**, Samples of S/L (left) and 2i/L (right) media with and without glutamine were analysed by GC-MS. Representative chromatograms of the total ion count reveal a clear glutamine (Q) peak in +Q media (grey) and no detectable glutamine in -Q media (red). m, minutes. **d**, Teratoma formation from ES cells grown in 2i/L medium without glutamine for 3 days. Representative images of haematoxylin and eosin staining reveal neural tissue (ectoderm), hepatocytes and pancreatic acinar cells (endoderm) and smooth muscle (mesoderm). Scale bar, 200 μ m. **e**, Growth curve of ESC-1 cells grown in glutamine-free 2i/L or 2i medium. **f**, Gene expression analysis confirms that EpiSCs, which represent post-implantation pluripotency, were generated from ESC-1 cells by culture with Fgf and activin A. Transcript levels were assessed by quantitative real-time polymerase chain reaction with reverse

transcription (qRT-PCR), normalized to *Gapdh* and expressed as a ratio of values of mouse ES cells cultured in 2i/L medium. **g**, Growth curve of EpiSCs cultured in serum-free epiblast medium (serum-free medium containing FGF and activin A (Fgf/ActA)) with or without glutamine. **h**, Growth curve of an iPSC line derived from fibroblasts using Oct3/4 (O), Klf4 (K) and Sox2 (S) cultured in glutamine-free S/L or 2i/L medium. **i**, Doubling time of ESC-1 cells cultured in glutamine-free 2i/L medium in the presence and absence of glutamine. **j**, Growth curve of ESC-V19 cells cultured in glutamine-free 2i/L medium in the presence or absence of 1 mM methylsulphoxide (MSO). **k**, ESC-V19 cells grown in glutamine-free S/L (left) or 2i/L (right) medium with or without 4 mM DM- α KG. For growth curve experiments, cells were seeded on day 0 in complete medium and then were changed to experimental medium on day 1 (indicated by red arrow). Data are presented as the mean \pm s.d. of triplicate wells from a representative experiment.



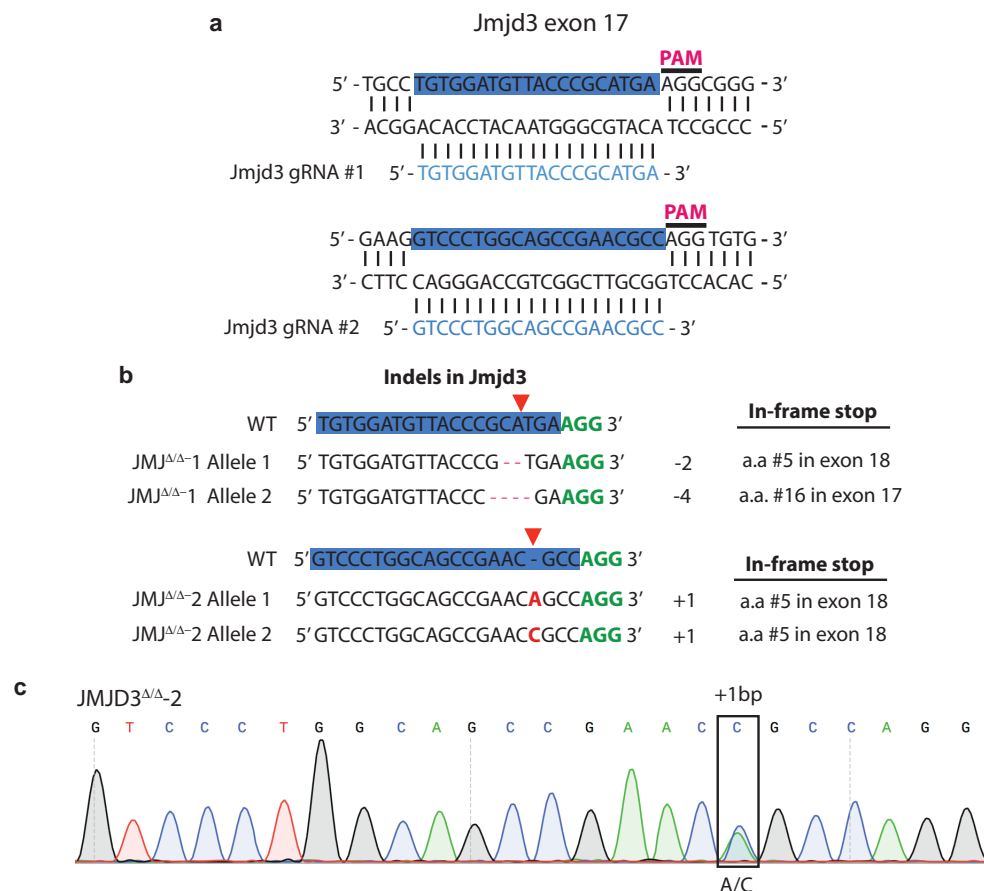
Extended Data Figure 2 | Mouse ES cells cultured with 2i demonstrate altered glucose and glutamine utilization. **a**, 2i enables glutamate synthesis from glucose-derived carbons. ESC-1 cells cultured in S/L, S/L/2i or 2i/L medium were incubated with medium containing $[\text{U-}^{13}\text{C}]$ glucose for 4 h and the fraction of glutamate containing glucose-derived carbons is shown. **b**, ESC-1 cells were cultured for 4 h in glutamine-free S/L or 2i/L medium containing $[\text{U-}^{13}\text{C}]$ glucose and the total amount of glutamate labelled by

glucose-derived carbons is shown. **c**, Incorporation of ^{14}C derived from $[\text{U-}^{14}\text{C}]$ glucose (^{14}C -glc) (left) or derived from $[\text{U-}^{14}\text{C}]$ glutamine (^{14}C -gln) (right) into total cellular protein after 48 h incubation. $P < 0.05$ for ^{14}C -glc, $P = 0.1$ for ^{14}C -gln, calculated by unpaired two-tailed Student's t -test. Data are presented as the mean \pm s.d. of triplicate wells (**a**, **b**) or \pm s.e.m. of quadruplicate wells (**c**) from a representative experiment.



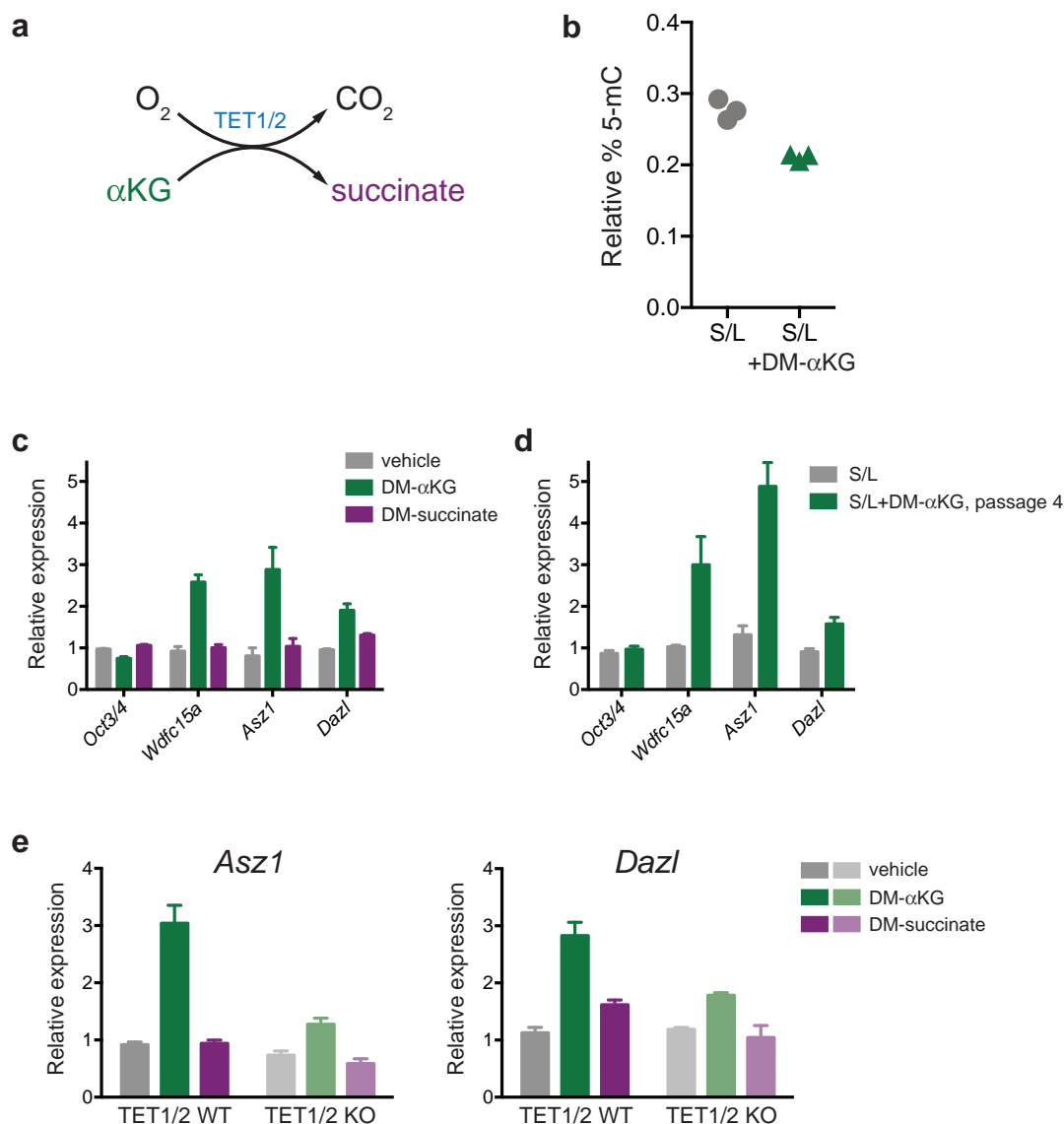
Extended Data Figure 3 | Regulation of histone methylation in 2i/LIF cells.
a, Western blot analysis of ESC-1 cells grown in glutamine-free 2i/L medium for 24 h with supplementation as indicated. **b**, **c**, H3K27me3 ChIP-qPCR of

ESC-1 cells cultured in S/L (**b**) or 2i/L (**c**) medium with or without 30 μ M GSK-J4 for 5 h. Data are presented as the mean \pm s.e.m. of triplicate samples.
 * $P < 0.05$ by unpaired Student's two-tailed t -test.



Extended Data Figure 4 | Generation of JMJD3 mutant cells. **a**, Schematic of targeting strategy for guide RNAs (gRNAs) to mouse *Jmjd3* exon 17. gRNA sequences are highlighted in blue. **b**, Representative sequences from two clones used in this study. Sanger sequencing revealed indels as shown in schematic. Red dashes, deleted bases; red bases, insertions. gRNA is highlighted in blue and

protospacer adjacent motif (PAM) sequences are identified in green. Predicted cut site is indicated by red triangle. Location of in-frame downstream stop is indicated on the right. **c**, An example chromatogram for clone JMJD3^{Δ/Δ}-2 showing single-base-pair insertions at the predicted Cas9 cleavage site.

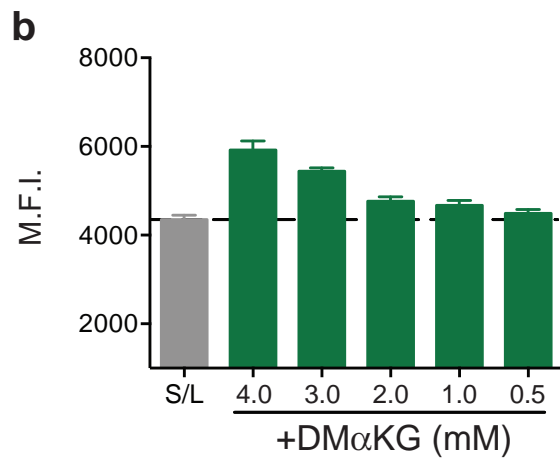
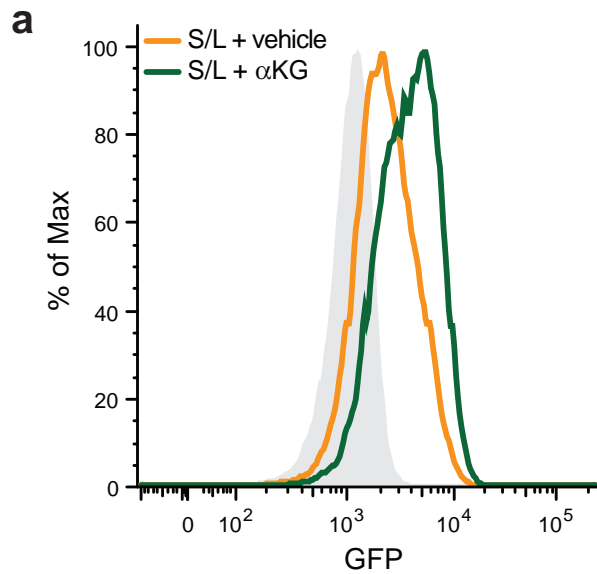


Extended Data Figure 5 | αKG increases Tet activity in mouse ES cells.

a, Simplified schematic of the reaction mechanism of Tet1/2 enzymes.

b, Relative per cent 5-methylcytosine (% 5-mC) in ESC-1 cells cultured in S/L medium with or without DM- αKG for 24 h. Each data point represents a sample from triplicate wells of a representative experiment. **c**, Gene expression in ESC-1 cells cultured with DM- αKG or DM-succinate for 3 days. **d**, Gene

expression in ESC-1 cells cultured in S/L medium with or without DM- αKG for four passages. **e**, Gene expression in wild-type or *Tet1/Tet2* double-knockout (KO) mouse ES cells cultured with DM- αKG or DM-succinate for 72 h. qRT-PCR data (**c–e**) was normalized to actin or *Gapdh* and samples were normalized to the control group. *Oct3/4* is not expected to change and is included as a control. Data are presented as the mean \pm s.e.m. of triplicate wells.



Extended Data Figure 6 | α KG increases Nanog expression.

a, Representative histogram of GFP intensity of Nanog-GFP cells treated with or without DM- α KG for 3 days. Grey represents background staining.

b, ESC-1 cells were cultured in S/L medium with DM- α KG for four passages and then switched to medium containing the indicated amounts of DM- α KG (0.5–4 mM) or vehicle control (S/L) for 3 days. GFP expression (M.F.I.) was determined by fluorescence-activated cell sorting (FACS). Data are presented as the mean \pm s.d. of triplicate wells from a representative experiment.

Deubiquitinase DUBA is a post-translational brake on interleukin-17 production in T cells

Sascha Rutz¹, Nobuhiko Kayagaki², Qui T. Phung³, Celine Eidenschenk¹, Rajkumar Noubade¹, Xiaoting Wang¹, Justin Lesch¹, Rongze Lu¹, Kim Newton², Oscar W. Huang⁴, Andrea G. Cochran⁴, Mark Vasser³, Benjamin P. Fauber⁵, Jason DeVoss¹, Joshua Webster⁶, Lauri Diehl⁶, Zora Modrusan⁷, Donald S. Kirkpatrick³, Jennie R. Lill³, Wenjun Ouyang¹§ & Vishva M. Dixit²§

T-helper type 17 (T_H17) cells that produce the cytokines interleukin-17A (IL-17A) and IL-17F are implicated in the pathogenesis of several autoimmune diseases^{1,2}. The differentiation of T_H17 cells is regulated by transcription factors such as RORγt^{3,4}, but post-translational mechanisms preventing the rampant production of pro-inflammatory IL-17A have received less attention. Here we show that the deubiquitylating enzyme DUBA is a negative regulator of IL-17A production in T cells. Mice with DUBA-deficient T cells developed exacerbated inflammation in the small intestine after challenge with anti-CD3 antibodies. DUBA interacted with the ubiquitin ligase UBR5, which suppressed DUBA abundance in naive T cells. DUBA accumulated in activated T cells and stabilized UBR5, which then ubiquitylated RORγt in response to TGF-β signalling. Our data identify DUBA as a cell-intrinsic suppressor of IL-17 production.

DUBA (also known as OTUD5) belongs to the ovarian tumour family of deubiquitylating enzymes⁵. DUBA limits type I interferon production in macrophages⁶, but its expression is not restricted to the myeloid lineage (Fig. 1a). CD4⁺ and CD8⁺ T cells expressed *Duba* (also called *Otud5*) messenger RNA, and transcript levels changed little after activation of the T-cell receptor (TCR) with anti-CD3 antibodies (Fig. 1b). DUBA protein, however, increased after TCR stimulation, and was the active enzyme phosphorylated on Ser 177 (ref. 7) (Fig. 1b). CD28 co-stimulation or IL-2 did not increase DUBA abundance further (Fig. 1c). Protein kinase C (PKC) signalling downstream of the TCR might post-translationally regulate the amount of DUBA because PKC activation by phorbol 12-myristate 13-acetate (PMA) was sufficient to increase DUBA expression in T cells (Fig. 1d). The proteasome inhibitor MG-132 also boosted the amount of DUBA in T cells (Fig. 1e), suggesting that DUBA is degraded by the proteasome in the absence of TCR signalling.

Mice with conditional *Duba* alleles (*Duba*^{fl/fl})⁷ and a CD4-Cre transgene⁸ were used to study the effect of deleting *Duba* in T cells (Extended Data Fig. 1a, b). *Duba*^{fl/fl} CD4-Cre mice aged 6 weeks exhibited normal thymus, spleen and mesenteric lymph node cellularity, but had two- to threefold fewer CD4⁺ CD8⁺ and CD4⁺ CD8⁺ thymocytes than *Duba*^{+/+} CD4-Cre mice (Extended Data Fig. 1c–e). CD8⁺ T cells were also reduced three- to fourfold in the spleen and lymph nodes, whereas the numbers of naive and memory CD4⁺ T cells in the periphery were normal (Extended Data Fig. 1f–h). These data indicate that DUBA deficiency perturbs the normal development of CD8⁺ T cells. *Duba*^{fl/fl} CD4-Cre mice also had fewer FOXP3⁺ T regulatory (T_{reg}) cells in the periphery, whereas numbers in the thymus were not significantly different from those seen in controls (Extended Data Fig. 1i, j).

When *Duba*^{+/+} and *Duba*^{-/-} CD4⁺ T cells were stimulated with PMA and ionomycin *ex vivo*, *Duba*^{-/-} T cells yielded significantly more IL-17A⁺ T_H17 cells, but normal numbers of IFN-γ⁺ T_H1 and IL-4⁺ T_H2 cells (Extended Data Fig. 2a). Accordingly, RORγt⁺ CD4⁺ T cells were more prevalent in *Duba*^{fl/fl} CD4-Cre mice (Extended Data Fig. 2b). Immunization

with ovalbumin (OVA) in complete Freund's adjuvant⁹ (CFA) further increased IL-17A⁺ CD4⁺ and RORγt⁺ CD4⁺ T-cell numbers in *Duba*^{fl/fl} CD4-Cre mice (Extended Data Fig. 2c, d). CD8⁺ T cells from naive *Duba*^{fl/fl} CD4-Cre mice also exhibited increased IL-17A production (Extended Data Fig. 2e). Most T_H17 cells in naive mice home to the small intestine¹⁰, and IL-17A⁺ CD4⁺ T cells were increased among intraepithelial lymphocytes from *Duba*^{fl/fl} CD4-Cre small intestine (Fig. 2a). We also examined the role of DUBA in type 3 innate lymphoid cells (ILCs),

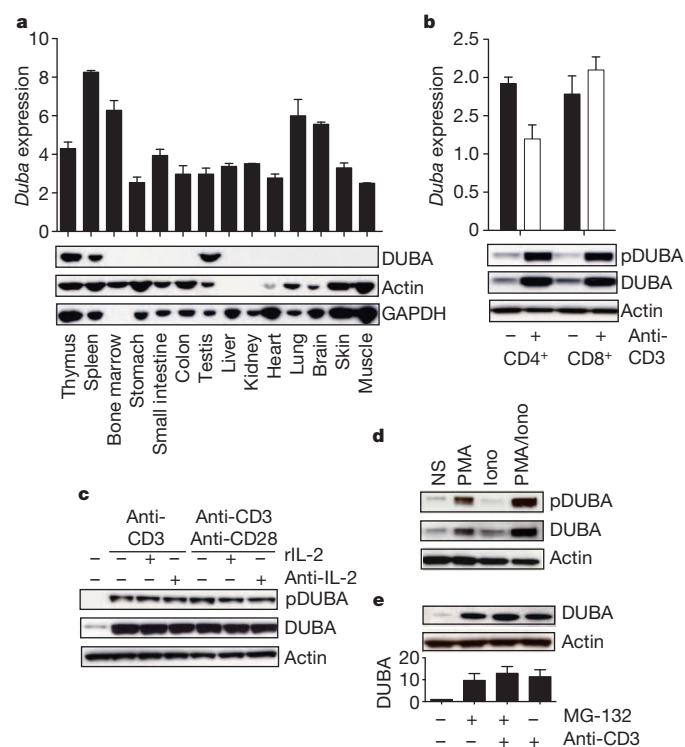


Figure 1 | Post-translational regulation of DUBA expression in T cells.

a, b, Relative expression of *Duba* mRNA (graphs) or DUBA protein (immunoblots) in mouse tissues (**a**) and CD4⁺ or CD8⁺ T cells (**b**). Where indicated, T cells were stimulated overnight with anti-CD3 antibodies. Error bars, s.e.m. of triplicate measurements. Results are representative of two independent experiments. pDUBA, DUBA phosphorylated on Ser 177. **c, d**, Immunoblots of CD4⁺ T cells. Where indicated, stimulation was overnight. Iono, ionomycin; NS, no stimulus; pDUBA, DUBA phosphorylated on Ser 177; rIL-2, recombinant IL-2. **e**, Immunoblots of CD4⁺ T cells stimulated for 6 h. Relative DUBA abundance from densitometry measurements. Error bars, s.e.m. of two independent experiments.

¹Department of Immunology, Genentech, 1 DNA Way, South San Francisco, California 94080, USA. ²Department of Physiological Chemistry, Genentech, 1 DNA Way, South San Francisco, California 94080, USA. ³Department of Protein Chemistry, Genentech, 1 DNA Way, South San Francisco, California 94080, USA. ⁴Department of Early Discovery Biochemistry, Genentech, 1 DNA Way, South San Francisco, California 94080, USA. ⁵Discovery Chemistry, Genentech, 1 DNA Way, South San Francisco, California 94080, USA. ⁶Department of Pathology, Genentech, 1 DNA Way, South San Francisco, California 94080, USA. ⁷Department of Molecular Biology, Genentech, 1 DNA Way, South San Francisco, California 94080, USA.

§These authors jointly supervised this work.

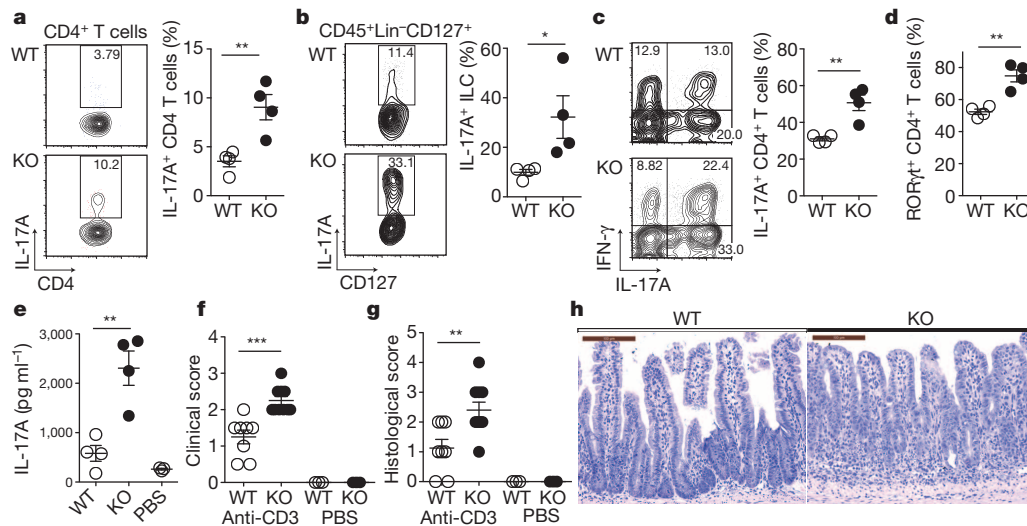


Figure 2 | DUBA limits IL-17A production. **a, b**, Graphs of the percentage of intraepithelial CD4⁺ T cells (**a**) or innate lymphoid cells (ILCs) (**b**) in the small intestine expressing IL-17A as determined by flow cytometry. Representative contour plots are shown. **c–h**, *Duba*^{fl/fl} CD4-Cre (knockout, KO) or *Duba*^{+/+} CD4-Cre (wild-type, WT) mice at 48 h after injection with anti-CD3 antibodies. **c, d**, Percentage of intraepithelial CD4⁺ T cells in the small intestine

expressing IL-17A (**c**) or RORγt (**d**). **e**, IL-17A in the serum. PBS indicates wild-type mice injected with vehicle alone. **f, g**, Clinical (**f**) and histological (**g**) scores as defined in the Methods. **h**, Small intestine stained with haematoxylin and eosin. Scale bars, 100 μm. Each circle in **a–g** denotes one mouse. Error bars, s.e.m. **P* < 0.05, ***P* < 0.01, ****P* < 0.001 (unpaired Student's *t*-test).

another source of IL-17A (ref. 11), using the pan-haematopoietic Vav-Cre deleter¹². IL-17A-producing type 3 ILCs were more prevalent in *Duba*^{fl/fl} Vav-Cre mice than in controls (Fig. 2b). Therefore, DUBA limits IL-17 production in several cell types.

Autoimmune diseases including rheumatoid arthritis, psoriasis and multiple sclerosis can feature increased production of IL-17A and increased numbers of T_H17 cells^{1,2,13–16}. After injection with anti-CD3 antibodies, which triggers T_H17-driven inflammation in the small intestine¹⁰, *Duba*^{fl/fl} CD4-Cre mice contained significantly more RORγt⁺ or IL-17A-expressing T_H17 cells in the small intestine, spleen and lymph nodes (Fig. 2c, d and Extended Data Fig. 3a, b), and more IL-17A in serum (Fig. 2e) when compared to controls. IFN-γ-expressing T cells and other serum cytokines were comparable to those in controls (Extended Data Fig. 3c–e). The enhanced T_H17 response in the *Duba*^{fl/fl} CD4-Cre mice coincided with more severe inflammation in the lamina propria of the small intestine and with increased epithelial cell proliferation and death (Fig. 2f–h and Extended Data Fig. 3f). These data suggest increased IL-17A production from *Duba*^{fl/fl} T cells exacerbates inflammation and tissue damage *in vivo*.

T_{reg} cells that make IL-17A exist in inflammatory environments in murine models and human diseases^{17–19}. Interestingly, more FOXP3⁺ T_{reg} cells were RORγt⁺ in the absence of DUBA (Extended Data Fig. 4a). In addition, *Duba*^{fl/fl} T_{reg} cells isolated *ex vivo* produced IL-17A after TCR stimulation, whereas their *Duba*^{+/+} counterparts did not (Extended Data Fig. 4b, c). Nonetheless, *Duba*^{fl/fl} T_{reg} cells exhibited normal suppressive function *in vitro* (Extended Data Fig. 4d), consistent with descriptions of IL-17-producing T_{reg} cells in other models^{19–21}. *Duba*^{fl/fl} T_{reg} cells also functioned normally in a T-cell transfer model of colitis^{22,23} (Extended Data Fig. 4e–h). Finally, *Duba*^{fl/fl} CD4-Cre mice showed no signs of autoimmunity after 1 year, despite having more IL-17-expressing CD4⁺ T cells (Extended Data Fig. 5).

Despite DUBA being expressed in all T-helper subsets (Extended Data Fig. 6a), DUBA deficiency increased IL-17A production only in T_H17 and induced T_{reg} (iT_{reg}) cells (Fig. 3a and Extended Data Fig. 6b). Moreover, IL-17A production was enhanced more in T_H17(TGF-β) cells derived with TGF-β and IL-6 than in T_H17(IL-1β) cells derived with IL-1β, IL-6 and IL-23 (ref. 24) (Fig. 3b). *Duba*^{fl/fl} T_H17(TGF-β) cells also expressed more IL-17F, IL-21 and IL-9 than their *Duba*^{+/+} counterparts, although levels of IL-22 and IL-10 were normal (Extended Data Fig. 6c). IL-17A production by *Duba*^{+/+} T_H17(TGF-β) cells *in vitro*

is transient unless IL-6 and TGF-β stimulation is maintained²². Of note, *Duba*^{fl/fl} T_H17(TGF-β) cells continued to secrete significant levels of IL-17 after secondary and tertiary TCR stimulation without TGF-β or IL-6 (Extended Data Fig. 6d). *Duba*^{fl/fl} CD8⁺ T cells cultured with TGF-β and IL-6 also expressed more IL-17A than controls, whereas T cytotoxic type 1 (T_C1) or T_C2 polarizing conditions elicited negligible IL-17A (Extended Data Fig. 6e).

Duba^{fl/fl} T_H1 and T_H2 cells exhibited normal expression of the transcription factors T-bet and GATA3, respectively (Extended Data Fig. 6f), whereas *Duba*^{fl/fl} T_H17(TGF-β) and iT_{reg} cells expressed more RORγt than *Duba*^{+/+} cells (Fig. 3c and Extended Data Fig. 6f). Other factors associated with T_H17 differentiation, including the closely related RORα, were expressed at equivalent levels in *Duba*^{+/+} and *Duba*^{fl/fl} T_H17 cells (Extended Data Fig. 7). Therefore, it was plausible that increased production of IL-17A in *Duba*^{fl/fl} T_H17 cells was a direct consequence of increased RORγt. Consistent with this idea, pharmacological inhibition of RORγt blocked IL-17A production in both *Duba*^{+/+} and *Duba*^{fl/fl} T_H17(TGF-β) cells (Fig. 3d).

Rorc mRNA encoding RORγt peaked after 24 h in both *Duba*^{+/+} and *Duba*^{fl/fl} T_H17(TGF-β) cultures, and was increased in the latter (Fig. 3e), whereas RORγt protein was markedly increased in the *Duba*^{fl/fl} cells (Fig. 3f). *Rorc* mRNA and RORγt protein levels were much lower in T_H17(IL-1β) cells than in T_H17(TGF-β) cells, with only a mild increase in RORγt in *Duba*^{fl/fl} T_H17(IL-1β) cells. We measured the half-life of RORγt in T_H17 cells after treatment with the translational inhibitor cycloheximide. Although RORγt was stable in *Duba*^{+/+} T_H17(IL-1β) cells for 6 h, it had a half-life of approximately 1 h in *Duba*^{+/+} T_H17(TGF-β) cells. Importantly, the half-life of RORγt extended beyond 6 h in *Duba*^{fl/fl} T_H17(TGF-β) cells (Fig. 3g). These data indicate that DUBA suppresses RORγt expression in response to TGF-β mostly by regulating RORγt protein stability and to a limited extent by regulating *Rorc* transcription.

DUBA loss did not compromise TGF-β signalling generally because expression of TGF-β receptors and phosphorylation of the TGF-β responsive transcription factors SMAD2 and SMAD3 appeared normal in *Duba*^{fl/fl} T cells (Extended Data Fig. 8). Comparing the transcriptomes of *Duba*^{+/+} and *Duba*^{fl/fl} T_H17 and iT_{reg} cells yielded few insights into how DUBA loss stabilized RORγt (data not shown). Proteins that were differentially ubiquitinated in *Duba*^{+/+} and *Duba*^{fl/fl} T_H17 cells by mass spectrometry²⁵ represented potential DUBA substrates. Sixteen proteins contained

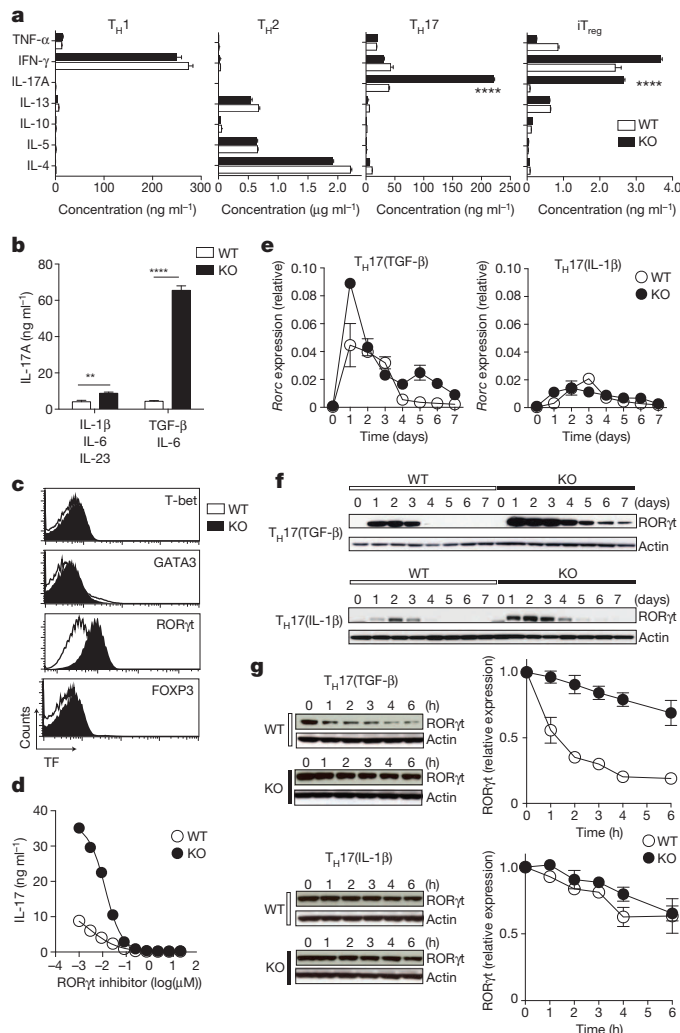


Figure 3 | DUBA deficiency in TH17 cells increases RORγt stability.

a, Cytokine production by T-helper subsets after 6 days of polarization *in vitro* and re-stimulation with anti-CD3/anti-CD28 antibodies for 48 h. Error bars, s.e.m. of triplicate measurements. Data are representative of two independent experiments. **b**, **d**, IL-17A production by TH17 cells. **c**, Flow cytometric analysis of T-bet, GATA3, RORγt and FOXP3 expression in TH17 cells. **e**, **f**, Expression of *Rorc* mRNA (**e**) and RORγt protein (**f**) in TH17 cells cultured for the times indicated. Error bars, s.e.m. of two independent measurements. Data are representative of two independent experiments. **g**, Immunoblots of RORγt in TH17 cells treated with cycloheximide for the times indicated. Graphs indicate relative RORγt abundance from densitometry measurements. Error bars, s.e.m. of three independent experiments. ** $P < 0.01$, **** $P < 0.0001$ (unpaired Student's *t*-test).

ubiquitinated peptides that were enriched more than 1.5-fold in *Duba*^{-/-} cells in two independent experiments, whereas an additional 11 proteins were enriched more than 1.5-fold in *Duba*^{+/+} cells (Fig. 4a and Extended Data Fig. 9). We also identified proteins co-immunoprecipitating with Flag-tagged DUBA (wild-type or a catalytically impaired Cys224Ser mutant) from retrovirally transduced TH17 cells (Fig. 4b).

PARP-1 was hyperubiquitinated in *Duba*^{-/-} cells and it interacted weakly with DUBA in co-immunoprecipitation experiments (Extended Data Fig. 10a). It had been reported to repress IL-17A production in TH17 cells, albeit by regulating TGF-β receptor expression and SMAD activity^{26–28}. Consistently, inhibition of PARP with veliparib increased IL-17A production in TH17(TGF-β) and iT_{reg} cells, and to a lesser extent in TH17(IL-1β) cells (Extended Data Fig. 10b). Veliparib also caused an early and transient increase in *Rorc* and *Il17a* mRNA levels in TH17 cells (Extended Data Fig. 10c), but it did not affect RORγt stability (Extended

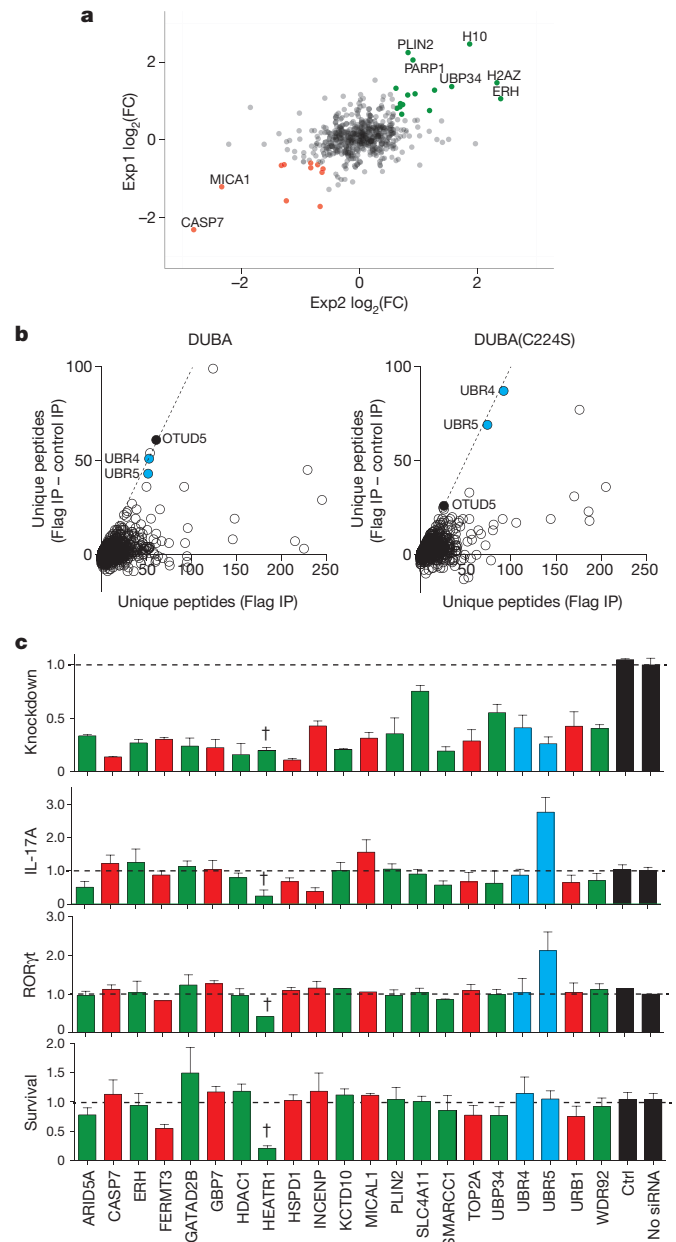


Figure 4 | Screens for substrates of DUBA. **a**, Correlation plot of proteins differentially ubiquitinated in two independent experiments (exp1 and exp2) comparing *Duba*^{+/+} and *Duba*^{-/-} TH17 cells after 4 days in culture. Green, ubiquitylation increased >1.5-fold in the *Duba*^{-/-} cells. Red, ubiquitylation decreased >1.5-fold in the *Duba*^{-/-} cells. FC, fold change. **b**, Proteins co-immunoprecipitated (IP) with Flag-tagged DUBA or mutant DUBA(Cys224Ser) expressed in TH17 cells. (Number of unique peptides in the experimental samples is plotted against the difference in the number of unique peptides in experimental samples and in control cells expressing green fluorescent protein (GFP).) **c**, Graphs indicate the efficiency of knockdown, IL-17A production, RORγt expression and cell survival after siRNA transfection of TH17 cells. Results are plotted relative to control (ctrl) transfections. Dagger symbol indicates impaired viability. Error bars, s.e.m. of two independent experiments.

Data Fig. 10d). DUBA loss did not affect PARP-1 abundance either (Extended Data Fig. 10e). Collectively, these data exclude PARP-1 as the critical substrate of DUBA in mediating RORγt stability.

Short interfering RNA (siRNA)-mediated knockdown of candidate DUBA substrates and interacting proteins identified UBR box E3 ligase UBR5 (also known as EDD)²⁹ as a crucial regulator of both IL-17A production and RORγt expression (Fig. 4c). Immunoblotting for UBR5 after

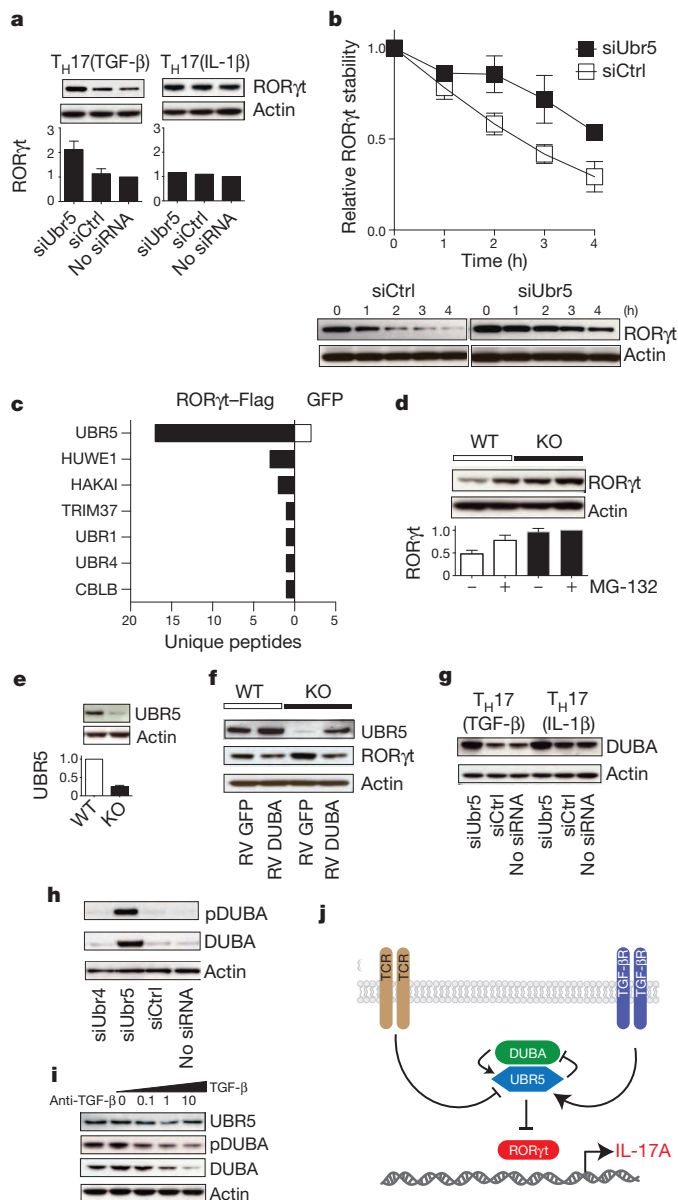


Figure 5 | Regulation of ROR γ t stability by UBR5 in TH17 cells.

a, Immunoblots of ROR γ t in wild-type TH17 cells transfected with the siRNA indicated. siUbr5, siRNA targeting *Ubr5*; siCtrl, non-targeting control siRNA. Graphs in **a**, **b** and **d** indicate relative ROR γ t abundance from densitometry measurements. Error bars, s.e.m. of two independent experiments. **b**, Immunoblots of TH17 cells after siRNA transfection, 3 days of culture, and then treatment with cycloheximide for the times indicated. Error bars, s.e.m. of three independent experiments. **c**, Mouse E3 ligases co-immunoprecipitated with Flag-tagged ROR γ t from reconstituted *Rorc*^{-/-} TH17 cells. Graph indicates the number of unique peptides in experimental samples (black) or control cells expressing GFP (white). **d**, Immunoblots of TH17 cells after 3 days of culture and addition of MG-132 for 4 h. Error bars, s.e.m. of two independent experiments. **e**, Expression of UBR5 protein in CD4⁺ T cells (relative UBR5 abundance from densitometry measurements). Error bars, s.e.m. of three independent experiments. **f**, Immunoblots of T cells transduced with DUBA- or GFP-expressing retroviruses (RV). **g**, Immunoblots of TH17 cells after siRNA transfection. **h**, Immunoblots of non-activated CD4⁺ T cells 24 h after siRNA transfection. **i**, Immunoblots of CD4⁺ T cells cultured with TGF- β or a neutralizing antibody to TGF- β . **j**, Model for regulation of ROR γ t expression and IL-17 production by DUBA and UBR5. TGF- β R, TGF- β receptor.

affinity purification of Flag–DUBA confirmed the interaction between DUBA and UBR5 (Extended Data Fig. 10f). siRNA knockdown of UBR5 (Extended Data Fig. 10g) mimicked DUBA deficiency by

increasing IL-17A production and ROR γ t expression in TH17(TGF- β) but not TH17(IL-1 β) cells (Fig. 5a and Extended Data Fig. 10h). In addition, UBR5 knockdown increased ROR γ t stability (Fig. 5b), suggesting that ubiquitylation of ROR γ t by UBR5 caused its proteasomal degradation. Indeed, endogenous UBR5 co-immunoprecipitated with Flag-tagged ROR γ t expressed ectopically in *Rorc*-deficient T cells (Fig. 5c), consistent with UBR5 targeting ROR γ t directly.

We speculated that DUBA and UBR5 deficiency gave similar phenotypes because DUBA suppresses expression of ROR γ t indirectly by stabilizing UBR5. Consistent with DUBA being required for proteasomal degradation of ROR γ t, MG-132 increased the amount of ROR γ t in *Duba*^{+/+} TH17(TGF- β) cells to that seen in *Duba*^{-/-} TH17(TGF- β) cells, but had no effect on ROR γ t abundance in *Duba*^{-/-} cells (Fig. 5d). Interestingly, *Duba*^{-/-} T cells contained less UBR5 protein than *Duba*^{+/+} T cells, despite expressing *Ubr5* mRNA normally (Fig. 5e and Extended Data Fig. 10i). Ectopic expression of DUBA in *Duba*^{-/-} TH17(TGF- β) cells was sufficient to restore UBR5 and ROR γ t levels to that seen in *Duba*^{+/+} cells (Fig. 5f). Unexpectedly, UBR5, in turn, appeared to regulate expression of DUBA because cells contained more DUBA after UBR5 knockdown (Fig. 5g). We propose that UBR5 ligase activity towards DUBA causes DUBA degradation in resting T cells, which express as much *Duba* mRNA as activated T cells (Fig. 1b). Indeed, siRNA knockdown of UBR5, but not UBR4, increased the amount of DUBA in resting T cells (Fig. 5h).

Given that expression of UBR5 changed little after T-cell activation (Extended Data Fig. 10j), trace amounts of DUBA seem sufficient to stabilize UBR5 in resting T cells (Figs 1b and 5e). We propose that TCR signalling inactivates UBR5 by an unknown mechanism, causing DUBA to accumulate. UBR5 activity is at least partially restored by the TGF- β pathway to limit the abundance of ROR γ t and its other substrates. Thus, titration of TGF- β during T-cell activation did not affect UBR5 expression, but altered the amount of DUBA in a concentration-dependent manner (Fig. 5i). Our model (Fig. 5j) also explains why DUBA was slightly less abundant in TH17 cells than in other T-cell subsets (Extended Data Fig. 6a), or in TH17(TGF- β) cells compared to TH17(IL-1 β) cells (Fig. 5g). This post-translational regulation adds an additional layer of complexity to the transcriptional networks known to drive TH17 development and function.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 12 February; accepted 16 October 2014.

Published online 3 December 2014.

- Hemdan, N. Y. A. *et al.* Interleukin-17-producing T helper cells in autoimmunity. *Autoimmun. Rev.* **9**, 785–792 (2010).
- Miossec, P. & Kolls, J. K. Targeting IL-17 and TH17 cells in chronic inflammation. *Nature Rev. Drug Discov.* **11**, 763–776 (2012).
- Yosef, N. *et al.* Dynamic regulatory network controlling TH17 cell differentiation. *Nature* **496**, 461–468 (2013).
- Ciofani, M. *et al.* A validated regulatory network for Th17 cell specification. *Cell* **151**, 289–303 (2012).
- Komander, D., Clague, M. J. & Urbé, S. Breaking the chains: structure and function of the deubiquitinases. *Nature Rev. Mol. Cell Biol.* **10**, 550–563 (2009).
- Kayagaki, N. *et al.* DUBA: a deubiquitinase that regulates type I interferon production. *Science* **318**, 1628–1632 (2007).
- Huang, O. W. *et al.* Phosphorylation-dependent activity of the deubiquitinase DUBA. *Nature Struct. Mol. Biol.* **19**, 171–175 (2012).
- Lee, P. P. *et al.* A critical role for Dnm1 and DNA methylation in T cell development, function, and survival. *Immunity* **15**, 763–774 (2001).
- Liang, S. C. *et al.* Interleukin (IL)-22 and IL-17 are coexpressed by Th17 cells and cooperatively enhance expression of antimicrobial peptides. *J. Exp. Med.* **203**, 2271–2279 (2006).
- Espigues, E. *et al.* Control of TH17 cells occurs in the small intestine. *Nature* **475**, 514–518 (2011).
- Spits, H. & Cupedo, T. Innate lymphoid cells: emerging insights in development, lineage relationships, and function. *Annu. Rev. Immunol.* **30**, 647–675 (2012).
- Georgiadis, P. *et al.* VavCre transgenic mice: a tool for mutagenesis in hematopoietic and endothelial lineages. *Genesis* **34**, 251–256 (2002).
- Chabaud, M., Fossiez, F., Taupin, J. L. & Miossec, P. Enhancing effect of IL-17 on IL-1-induced IL-6 and leukemia inhibitory factor production by rheumatoid arthritis synoviocytes and its regulation by Th2 cytokines. *J. Immunol.* **161**, 409–414 (1998).

14. Chabaud, M. *et al.* Human interleukin-17: A T cell-derived proinflammatory cytokine produced by the rheumatoid synovium. *Arthritis Rheum.* **42**, 963–970 (1999).
15. Lock, C. *et al.* Gene-microarray analysis of multiple sclerosis lesions yields new targets validated in autoimmune encephalomyelitis. *Nature Med.* **8**, 500–508 (2002).
16. Matusevicius, D. *et al.* Interleukin-17 mRNA expression in blood and CSF mononuclear cells is augmented in multiple sclerosis. *Mult. Scler.* **5**, 101–104 (1999).
17. Yang, X. O. *et al.* Molecular antagonism and plasticity of regulatory and inflammatory T cell programs. *Immunity* **29**, 44–56 (2008).
18. Voo, K. S. *et al.* Identification of IL-17-producing FOXP3⁺ regulatory T cells in humans. *Proc. Natl Acad. Sci. USA* **106**, 4793–4798 (2009).
19. Kryczek, I. *et al.* IL-17⁺ regulatory T cells in the microenvironments of chronic inflammation and cancer. *J. Immunol.* **186**, 4388–4395 (2011).
20. Beriou, G. *et al.* IL-17-producing human peripheral regulatory T cells retain suppressive function. *Blood* **113**, 4240–4249 (2009).
21. Ma, C. & Dong, X. Colorectal cancer-derived Foxp3⁺ IL-17⁺ T cells suppress tumour-specific CD8⁺ T cells. *Scand. J. Immunol.* **74**, 47–51 (2011).
22. Powrie, F., Leach, M. W., Mauze, S., Caddle, L. B. & Coffman, R. L. Phenotypically distinct subsets of CD4⁺ T cells induce or protect from chronic intestinal inflammation in C. B-17 *scid* mice. *Int. Immunol.* **5**, 1461–1471 (1993).
23. Mottet, C., Uhlig, H. H. & Powrie, F. Cutting edge: cure of colitis by CD4⁺CD25⁺ regulatory T cells. *J. Immunol.* **170**, 3939–3943 (2003).
24. Ghoreschi, K. *et al.* Generation of pathogenic T_H17 cells in the absence of TGF- β signalling. *Nature* **467**, 967–971 (2010).
25. Kim, W. *et al.* Systematic and quantitative assessment of the ubiquitin-modified proteome. *Mol. Cell* **44**, 325–340 (2011).
26. Zhang, P. *et al.* PARP-1 regulates expression of TGF- β receptors in T cells. *Blood* **122**, 2224–2232 (2013).
27. Sterling, J. A., Wu, L. & Banerji, S. S. PARP regulates TGF- β receptor type II expression in estrogen receptor-positive breast cancer cell lines. *Anticancer Res.* **26**, 1893–1901 (2006).
28. Lönn, P. *et al.* PARP-1 attenuates Smad-mediated transcription. *Mol. Cell* **40**, 521–532 (2010).
29. Callaghan, M. J. *et al.* Identification of a human HECT family protein with homology to the *Drosophila* tumor suppressor gene *hyperplastic discs*. *Oncogene* **17**, 3479–3491 (1998).

Acknowledgements We thank A. C. Chan for discussions, Q. Song for statistical analysis, and C. Bakalarski for informatics support.

Author Contributions S.R., N.K., Q.T.P., C.E., R.N., X.W., J.L., R.L., O.W.H., J.D. and Z.M. performed experiments designed and analysed by S.R., D.S.K., J.R.L., A.G.C., W.O. and V.M.D.; J.W. and L.D. performed histopathology; M.V. synthesized siRNAs; S.R., W.O. and K.N. wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to W.O. (ouyang@gene.com) or V.M.D. (dixit@gene.com).

METHODS

Mice. C57BL/6 mice (Jackson Laboratories), *Rorc*^{GFP/GFP} mice (Jackson Laboratories) and *Rag2*^{-/-} mice (Taconic) were housed under specific pathogen-free (SPF) conditions and used at 8–12 weeks of age. *Duba*^{fl/fl} mice⁷ (Extended Data Fig. 1a) were crossed to CD4-Cre mice to generate T-cell-specific knockout mice (*Duba*^{fl/fl} CD4-Cre). *Duba*^{fl/fl} mice⁷ were crossed to Vav-Cre mice¹² to generate leukocyte-specific knockout mice (*Duba*^{fl/fl} Vav-Cre). The Genentech Institutional Animal Care and Use Committee approved all animal studies.

Plasmids. An amino-terminal Flag tag was added to full-length mouse *Duba* and cloned as a P2A-GFP construct into the pQCXIX retroviral expression vector (Clontech). A DUBA(Cys224Ser) mutant was generated by site-directed mutagenesis using the Quikchange Kit (Agilent Technologies) according to the manufacturer's protocol. pQCXIX-GFP was used as a control vector. Full-length mouse *Rorc* was cloned into the pMSCV retroviral expression vector and a C-terminal Flag tag was added. pMSCV-GFP was used as a control vector.

T-cell purification. Naive CD4⁺ T cells were purified from spleen and mesenteric lymph node with CD4 microbeads (Miltenyi Biotec), followed by sorting for CD4⁺ CD45Rb^{hi} CD44^{low} CD25⁻ cells in a FACS Aria (BD Biosciences). Cells were stained with anti-CD4 (clone RM4-5), anti-CD45Rb (clone 16A), anti-CD44 (clone IM7) and anti-CD25 (clone PC61) antibodies (BD Biosciences). Alternatively, cells from the spleen and lymph node were stained with biotin-conjugated anti-CD25 (clone 7D4, BD Biosciences) and FITC-conjugated anti-CD4 (clone RM4-5, BD Biosciences). CD25⁺ cells were depleted using anti-biotin microbeads (Miltenyi Biotec). Next, CD4⁺ T cells were isolated using anti-FITC multisort beads (Miltenyi Biotec). After release of the beads, CD62L⁺ CD4⁺ T cells were purified using anti-CD62L microbeads (Miltenyi Biotec). Both procedures resulted in comparable purities of >97%. CD8 T cells were purified using the naive CD8 T-cell isolation kit (Miltenyi Biotec). T_{reg} cells were sorted by flow cytometry on the basis of their expression of CD4 and CD25.

T-cell cultures. Cells were cultured in DMEM with 10% FCS (HyClone), 2 mM L-glutamine, 1 mM sodium pyruvate, 0.1 mM non-essential amino acids, 55 μM β-mercaptoethanol, 100 U ml⁻¹ penicillin and 100 μg ml⁻¹ streptomycin (Invitrogen). Plates were coated overnight with 5 μg ml⁻¹ anti-CD3 antibody (clone 145-2C11, BD Biosciences) in PBS. Anti-CD28 (clone 37.51, BD Biosciences) (1 μg ml⁻¹) was added to naive T-cell cultures under T_{H0}, T_{H1}, T_{H2}, T_{H17} or iT_{reg} conditions. T_{H0} polarization: 10 μg ml⁻¹ anti-IL-4 (clone 11B11, in-house), 10 μg ml⁻¹ anti-IFN-γ (clone H22, in-house); T_{H1} polarization: 10 μg ml⁻¹ anti-IL-4, 10 ng ml⁻¹ rIL-12 (R&D Systems); T_{H2} polarization: 10 μg ml⁻¹ anti-IFN-γ, 20 ng ml⁻¹ rIL-4 (R&D Systems); T_{H17} polarization: anti-IL-4, 10 μg ml⁻¹ anti-IFN-γ, 20 ng ml⁻¹ rIL-6 (R&D Systems), 1 ng ml⁻¹ rhTGF-β (R&D Systems); iT_{reg} polarization: 10 μg ml⁻¹ anti-IL-4, 10 μg ml⁻¹ anti-IFN-γ, 5 ng ml⁻¹ rhTGF-β, 10 ng ml⁻¹ rIL-2 (R&D Systems). Where indicated, 10 μM MG-132 (EMD Millipore) or 10 μg ml⁻¹ cycloheximide (Sigma) were added. RORγt activation was blocked using a pharmacological inhibitor (Merck WO2012/106995) synthesized in-house. PARP-1 activity was blocked using 200 nM veliparib.

Retroviral transduction. Phoenix E cells were transiently transfected using the calcium phosphate method with retroviral expression plasmids together with the retroviral packaging plasmid pEco. Viral supernatant was collected after 48 h, filtered through a 0.45-μm filter, supplied with 10 mM HEPES, pH 7.2, and 10 μg ml⁻¹ polybrene (EMD Millipore), and added to T cells that had been stimulated for 24 h. T cells (2.5 × 10⁶) were incubated in 12-well plates with 1.5 ml of viral supernatant and centrifuged for 75 min at 700g at 32 °C. Thereafter, viral supernatant was replaced with conditioned culture medium.

siRNA knockdown. Chemically modified siRNAs³⁰ were synthesized in-house (dX, deoxy nucleotide; mX, methoxy nucleotide): siCtrl: sense 5'-GmGmAGCGCA CCAUCUUCdCdAdAmAmTmT-3', antisense 5'-dTUGAGAAGAUGGUGCGC UmCmC-3'; siUbr2: sense 5'-mAmACCUGAUGACCGUUAUdCdAdTmUmA mA-3', antisense 5'-dAAUGAUAAACGGUACAGGmUmU-3'; siUbr4: sense 5'-mUmGGCGGCUUCCUGAAAdCdAdTmAmCmC-3', antisense 5'-dTAAUG UUCAGGGAAGCCGmCmA-3'; siUbr5: sense 5'-mCmAGCAAAGAGGGA GAGUdCdAdAmAmCmC-3', antisense 5'-dTUUGACUCUCCUCUUUGCmU mG-3'; *Arid5a*: sense 5'-mGmACCACUGAGAAGCUGAdAdGdAmAmGmG-3', antisense 5'-dTUCUACAGCUUCAGUGGmUmC-3'; *Casp7*: sense 5'-mGmA GGACUGAUUUACGGdGdAdAmAmGmA-3', antisense 5'-dTUUCGCGUAAA UCAGGUCCmUmC-3'; *Erlh*: sense 5'-mUmCUGGGAAAGUCUCAGUdGdAdAm AmAmA-3', antisense 5'-dTUUCACUGAGACUUUCCAmGmA-3'; *Fermt3*: sense 5'-mAmGGAGAAGAAGAAGdGdAdGdAmAmG-3', antisense 5'-dTUCU CUCUUCUUCUUCmCmU-3'; *Gatad2b*: sense 5'-mGmCGAGAUGAUGUU CUGdCdAdAmAmA-3', antisense 5'-dTUUGCCAGAACAUAUCUAmCmG C-3'; *Gbp7*: sense 5'-mAmAGGAAGAACAUCUGAdGdAdGmAmGmC-3', anti- sense 5'-dTUCUCAGGAUGUUCUUCmUmU-3'; *Heatr1*: sense 5'-mUmAGA GAAAAUCACUAGAdGdAdAmAmUmG-3', antisense 5'-dTUUCUCUAGUGAU UUCUCmUmA-3'; *Hspd1*: sense 5'-mUmGUGUACAAAGUAGAGAdAdGdT

mAmUmC-3', antisense 5'-dTACUUCUCUACUUUGUACAmCmA-3'; *Incpn*: sense 5'-mUmCAGAAGAACUGGAGdAdGdAmAmGmA-3', antisense 5'-dT UCUCUCCAGAUCCUUCmGmA-3'; *Kctd10*: sense 5'-mUmUGUAAACAGC CAGAUGdAdTdTmAmUmU-3', antisense 5'-dTAAUACUCUGGCUUGUUAACm AmA-3'; *Plin2*: sense 5'-mCmUGGUCACGCCCAAGGdGdTdTmAmCmC-3', antisense 5'-dTAAACCUUGGGCGUUGACmAmG-3'; *Slc4a11*: sense 5'-mUm CGCACAGAGGAAGAAUdTdTmAmAmGmG-3', antisense 5'-dTUGAAUUCU UCCUCUGUGCmGmA-3'; *Top2a*: sense 5'-mUmCAAACAGACGUGGAUGd GdAdTmAmAmC-3', antisense 5'-dTAAUCCACGUCUGUUUmGmA-3'; *Urb1*: sense 5'-mAmGGGGAAGCAUCAUGdCdAdTmAmUmA-3', antisense 5'-dT AUGCCAUGAUGCUUUCmCmU-3'; *Wdr92*: sense 5'-mCmAGCUUG GAAUUGACAdGdAdAmAmGmG-3', antisense 5'-dTUUCUGUCAAUUC CAAGCmUmG-3'; *Ubr4*: sense 5'-mUmCAGAAUUGUCUUGAGAdTdTm AmCmG-3', antisense 5'-dTAGAUCUCAAGACAUUUCUmGmA-3'; *Ubr5*: sense 5'-mUmUAGAGAAAGCUAGAGCdAdAdAmAmAmA-3', antisense 5'-dTUUU GCUCUAGCUUUCUCUmAmA-3'; *Mical1*: sense 5'-mCmUGUGUGAACUC UGUGdGdAdAmAmCmA-3', antisense 5'-dTUUCACAGAGUUCACACm AmG-3'; *Usp34*: sense 5'-mGmAGGAAGAAGAUUGAdGdAdGdAmAmGmA-3', antisense 5'-dTUCUUAUCAUCUUCUUCmUmC-3'; *Hdac1*: sense 5'-mUm AAAACAGAGGAUGAGAdAdGdAmAmGmA-3', antisense 5'-dTUUUCUCAU CCUCUGUUUmUmA-3'; *Smarrc1*: sense 5'-mAmUGGAUGAAUGAAGAGGdA dTdTmAmUmG-3', antisense 5'-dTAAUCCUUCUUAUUAUCCmAmU-3'.

Naive T cells were transfected with 600 pmol siRNA by nucleofection (Lonza) under the following conditions: 3 × 10⁶ cells, mouse T-cell nucleofection solution, program X-001. Cells were transferred directly into DMEM medium with 10% FCS, 2 mM L-glutamine, 1 mM sodium pyruvate, 0.1 mM non-essential amino acids, 55 μM β-mercaptoethanol, 100 U ml⁻¹ penicillin and 100 μg ml⁻¹ streptomycin, rested for 2 h and then cultured and stimulated as indicated.

In vitro suppression assay. T_{reg} cell *in vitro* suppression assays were performed as described³¹. In brief, CD4⁺ T cells (40,000 per well) depleted of CD25⁺ T_{reg} cells were labelled with CFSE (Invitrogen) according to the manufacturer's protocol, and co-cultured with CD4⁺ CD25⁺ T_{reg} cells at different ratios (1:2 to 1:64). T cells were stimulated with irradiated antigen-presenting cells APC (enriched for MHCII positive cells by MACS, 80,000 per well) and 1 μg ml⁻¹ anti-CD3. After 3 days, proliferation was analysed by flow cytometry.

Re-stimulation and intracellular staining. T cells were stimulated for 4 h with 10 ng ml⁻¹ PMA (Sigma) and 1 μg ml⁻¹ ionomycin (Invitrogen). Brefeldin A (eBioscience) was added at 5 μg ml⁻¹. After fixation with 2% paraformaldehyde, the cells were permeabilized in CytoPerm buffer (BD Biosciences) and stained intracellularly for IL-17 (clone TC11-18H10, BD Biosciences), IL-4 (clone 11B11, eBioscience), IFN-γ (clone XMG1.2 BD Biosciences) and CD4 (clone RM4-5, BD Biosciences) or CD8 (clone 53-6.7, BD Biosciences) for 20 min at 20 °C.

For intracellular staining of transcription factors, cells were fixed overnight and permeabilized using the FOXP3 staining kit (eBioscience) and stained for FOXP3 (clone FJK-16s, eBioscience), RORγt (clone B2D, eBioscience), T-bet (clone eBio4B10, eBioscience) or GATA3 (clone TWAJ, eBioscience). Samples were acquired on a FACS Calibur flow cytometer (BD Biosciences) and data analysis was conducted using FlowJo (Tree Star). Intracellular staining for FOXP3 (clone FJK-16s, eBioscience) and RORγt (clone B2D, eBioscience) was performed with overnight fixation using the anti-Mouse/Rat Foxp3 Staining Set (eBioscience).

ELISA. IL-17A, IL-17F, IL-21, IL-9 and IL-10 production was analysed from supernatants taken on day 3 of culture using Ready-Set-Go kits (eBioscience). An IL-22 ELISA was developed in-house and performed as previously described³². In some instances, supernatants were analysed for cytokine production using 23-plex Luminex (BioRad).

RNA isolation and real-time RT-PCR. RNA was isolated using an RNeasy Mini Kit (Qiagen). Mouse tissue total RNA was purchased from Zyagen. RNA samples were analysed by real-time RT-PCR with TaqMan One-Step RT-PCR Master Mix reagents (Applied Biosystems) with the primers and probes indicated below. Results were normalized to those of the control housekeeping gene *Rpl19* (encoding ribosomal protein L19) and are reported as 2^{-ΔCt}.

Rpl19 forward 5'-GCATCTCATGGAGCACAT-3', reverse 5'-CTGGTCAGC CAGGAGCTT-3', probe 5'-CTTGCGGGCCTTGTCTGCCTT-3'; *Duba* forward 5'-AGTCCGGAACGTGAAGAGGT-3', reverse 5'-TCAAACAGTGTCTCTG CTG-3', probe 5'-AGCGGGCTACAACAGTGAAG-3'; *Ubr2* forward 5'-AGGC AAACAAGCCTTCTCAC-3', reverse 5'-AAAACACAGGTGGGGTCAAC-3', probe 5'-TGGCCGAGTGTAAAGTGGGG-3'; *Ubr4* forward 5'-GCACTGTCCA CGCATCAT-3', reverse 5'-GCAGGAGGAAAACCTAATG-3', probe 5'-CA TCGGTGGAGCTGCCTG-3'; *Ubr5* forward 5'-TGATGAGGATGGAGATG ACG-3', reverse 5'-GTGGGCAGAGTGAATGTGAC-3', probe 5'-GCCAGTGAA TCTTACTGCGCGG-3'; *Il17a* forward 5'-GCTCCAGAAGGCCCTCAGA-3', reverse 5'-CTTTCCTCCGATTGACA-3', probe 5'-ACCTCAACCGTTCCA CGTCAC-3'; *Il17f* forward 5'-TGAAACAGCCATGGTCAAGTCT-3', reverse

5'-CGAGCTGCTACCTCCCTCAG-3', probe 5'-TGCTACTGTTGATGTTGGGA
CTTGCCA-3'; *Il9* forward 5'-ATCAGTGTCCGTCCTTTTC-3', reverse 5'-GT
CTTCATGGTCGGCTTTTC-3', probe 5'-CCATGCAACCAGACCATGGCA-3';
Il21 forward 5'-AAGATTCTGAGGATCCGAGAAG-3', reverse 5'-GCATTCG
TGAGCGTCTATAGTGTG-3', probe 5'-TTCCCGAGGACTGAGGAGACGCC-3';
Il22 forward 5'-TCCGAGGAGTCACTGCTAA-3', reverse 5'-AGAACGTCTTC
CAGGGTGAA-3', probe 5'-TGAGCACCTGCTTCATCAGTAGCA-3'; *Il10*
forward 5'-CAGCCGGGAAGACAATAACT-3', reverse 5'-ATGTTGTCCAGCT
GGTCTT-3', probe 5'-GCTGCGGACTGCCTTCAGCC-3'; *Rorc* forward 5'-CGA
GATGCTGTCAAGTTTGG-3', reverse 5'-CACTTGTCTCTGTTGCTGCT-3', probe
5'-CCCTCTGCTTCTTGACATTCGG-3'; *Foxp3* forward 5'-CTGGGCTTCT
GGGTATGTC-3', reverse 5'-GAGCCCATGGCAGAGTT-3', probe 5'-TTCC
CTCCATCCACCTAAGCAGC-3'; *Stat3* forward 5'-CCCCGCACTTATAGT
TCATG-3', reverse 5'-CCCTCTGCTGAGGGCTC-3', probe 5'-TGCAAGTTT
GGAAATAACGGTGAAGGTGC-3'; *Batf* forward 5'-TGGCAACAGGACTC
ATCTG-3', reverse 5'-TGTCGGCTTCTGTGTCTGT-3', probe 5'-CGCTGCCC
AGAAGAGCCGAC-3'; *Irf4* forward 5'-GCCCAACAAGCTAGAAAG-3', reverse
5'-TCTCTGAGGGTCTGAAACT-3', probe 5'-AGTTTCTATCAGAGCTGCA
AGTGTGTGCTCA-3'; *Maf* forward 5'-CCCGTAACCTGCTAGCTTGG-3', reverse
5'-TTGTGCAAGTCCAGGGTATG-3', probe 5'-TTTACACCGCTGCCCA
CA-3'; *Ahr* forward 5'-CCTGGAATTCGAACCAAAA-3', reverse 5'-AACTGG
TACCCCGATCCTCT-3', probe 5'-TGGTTGTGATGCCAAAGGGCA-3'; *Irf8*
forward 5'-ACAATCAGGAGGTGGATGCT-3', reverse 5'-GGCTGGTTCAGC
TTTGTCTC-3', probe 5'-TCCATCTTCAAGGCTGGGCA-3'; *Rum1* forward
5'-CCAGCCTCTGCGAAGCTT-3', reverse 5'-GACGCGAGTAGGGAAC
TG-3', probe 5'-CAACGGCTCCGACCTGACC-3'; *Socs3* forward 5'-GATTT
CGCTTCGGGACTACT-3', reverse 5'-GGAACCTGCTGTGGGTGA-3', probe
5'-AAGCAGCTGCAGCCACCGC-3'; *Hif1a* forward 5'-TGCTCATCAGTTGC
CACTTC-3', reverse 5'-CCATCTGTGCTTCATCTCA-3', probe 5'-TGGATGC
CGGTGCTGACAGATG-3'; *Tgfb1* forward 5'-CAGACTGGGACTTGCT
GTG-3', reverse 5'-TAGAATTCAGGGGCCATGT-3', probe 5'-TTGCTCCA
AACCACAGAGTAGGCACT-3'; *Tgfb2* forward 5'-AACATGGAAGAGTGC
AACGA-3', reverse 5'-TGACACCCGTCATTTGGATA-3', probe 5'-CCACCA
CGAGTCCCAGCTG-3'; *Ubr4* forward 5'-GCACTGTCCACGCACACTAC-3',
reverse 5'-GCAGGAGGGAACAATG-3', probe 5'-CAGTCGGTGGCAG
TGCTG-3'; *Ubr5* forward 5'-TGATGAGGATGAGATGACG-3', reverse 5'-GTG
GGCAGAGTGAATGTGAG-3', probe 5'-GCCAGTGAATCTTACCTGCCGGG-3'.

Taqman sets from Life Technologies: *Arid5a* Mm00524454, *Casp7* Mm00432324,
Erh Mm01302670, *Fermt3* Mm00464182, *Gatad2b* Mm00522590, *Gbp7* Mm00523797,
Hdac1 Mm02391771, *Heat1* Mm00522863, *Hsp1* Mm00849835, *Incpm* Mm01198109,
Kctd10 Mm00525311, *Mical1* Mm00506780, *Plin2* Mm00475794, *Slc4a11* Mm01329596,
Smarc1 Mm00486224, *Top2a* Mm01296339, *Ubp34* Mm01231150, *Urb1* Mm01209640
and *Wdr92* Mm00724835.

Flag immunoprecipitations. T cells were washed with ice-cold PBS, re-suspended in lysis buffer (50 mM Tris-HCl, pH 7.4, 150 mM NaCl, 1% Triton X-100, 1 mM EDTA, Halt protease and phosphatase inhibitor (Thermo Scientific)), and incubated on ice for 30 min. Soluble lysates were generated by centrifugation at 16,000g for 10 min at 4 °C, and then incubated with anti-Flag M2 affinity gel (Sigma) for 4 h at 4 °C with rotation. Resin was washed four times for 5 min at 4 °C with lysis buffer and eluted in two consecutive steps in 0.2 mg ml⁻¹ 3×Flag peptide in 50 mM Tris-HCl, pH 7.4, 150 mM NaCl at 4 °C with rotation.

Immunoblotting. Equal amounts of proteins, measured using BCA assay (Thermo Scientific), were separated by SDS-PAGE on 4–12% Bis-Tris gels (Invitrogen) and transferred onto nitrocellulose membranes using iBlot apparatus (Invitrogen). Membranes were blocked for 1 h at room temperature with 5% milk in TBST (10 mM Tris, pH 7.5, 150 mM NaCl, 0.1% (v/v) Tween-20) and probed with primary antibody in 5% milk overnight at 4 °C, washed extensively in TBST, and incubated at 20 °C for 1 h with anti-rabbit-HRP, anti-mouse-HRP (Cell Signaling), or anti-rat-HRP (Santa Cruz Biotechnology). Proteins were visualized with SuperSignal West Pico chemiluminescent substrate (Pierce) or SuperSignal West Femto chemiluminescent substrate (Pierce). Primary antibodies recognized DUBA (Genentech), phospho-DUBA S¹⁷⁷ (Genentech)⁶, RORγt (B2D, eBioscience), RORα (X-23, Santa Cruz Biotechnology), RXRβ (8715, Cell Signaling), RARα (2554, Cell Signaling), PPARδ (PA5-29678, Thermo Scientific), UBR5 (8755, Cell Signaling), PARP-1 (46D11, Cell Signaling), FOXF3 (eBio7979, eBiosciences), c-Maf (M-153, Santa Cruz Biotechnology), AhR (BML-SA210, Enzo Life Sciences), BATF (polyclonal, Genentech), IRF-4 (P173, Cell Signaling), IRF8 (D20D8, Cell Signaling), pSTAT3 (3E2, Cell Signaling), RUNX1 (D4A6, Cell Signaling), SOCS3 (2923, Cell Signaling), HIF1A (NB100-105, Novus Biologicals), TGF-βR1 (ABF17, Millipore), TGF-βR2 (11888S, Cell Signaling), SMAD2/3 (5678S, Cell Signaling), SMAD4 (9515S, Cell Signaling), pSMAD2 (3101S, Cell Signaling), pSMAD3 (9520S, Cell Signaling), actin (rabbit polyclonal, Sigma-Aldrich), and GAPDH (14C10, Cell Signaling). Immunoblots

were quantified using the software ImageJ (Rasband, W. S., ImageJ, US National Institutes of Health; <http://imagej.nih.gov/ij/>, 1997–2014.)

Mass spectrometry. Anti-Flag immunoprecipitates from primary T cells expressing Flag-DUBA or GFP as a control were reduced with 10 mM dithiothreitol (DTT) for 1 h at 37 °C, and alkylated with 15 mM iodoacetamide at room temperature in the dark for 30 min. Samples were then loaded onto a 4–12% Bis-Tris gel (Life Technologies). Entire gel lanes were excised and divided from top to bottom into 15–20 bands. The gel bands were de-stained with 50:50 methanol:50 mM ammonium bicarbonate in water then digested at 37 °C overnight with 0.02 μg μl⁻¹ trypsin (Promega) in 50 mM ammonium bicarbonate. The digested samples were injected onto a 100 μm inner diameter capillary column (NanoAcquity UPLC column, 100 μm × 100 mm, 1.7 μm, BEH130 C18, Waters Corp) and separated by capillary reverse phase chromatography on a NanoAcquity UPLC system (Waters Corp). Samples were loaded in 0.1% formic acid, 2% acetonitrile and 98% water and eluted with a gradient of 2–90% buffer B (in which buffer A is 0.1% formic acid, 2% acetonitrile and 98% water, and buffer B is 0.1% formic acid, 2% water and 98% acetonitrile) at 1 μl min⁻¹ with a total analysis time of 60 min. Peptides were eluted directly into an LTQ-Orbitrap Elite mass spectrometer (ThermoFisher) and ionized using an ADVANCE source (Michrom-Bruker) at a spray voltage of 1.2 kV. Mass spectral data were acquired using a method comprising of one full mass spectrometry scan (375–1,600 *m/z*) in the Orbitrap at 60,000 resolution *M/ΔM* (*M* = mass, *ΔM* = full-width of peak at half-maximum height) at *m/z* 400 followed by collision-induced of the top 15 most abundant ions detected in the full mass spectrometry scan in a cycle repeated throughout the liquid chromatography gradient in the linear ion trap.

Tandem mass spectral results were submitted for database searching using the Mascot search algorithm ver 2.3.02 (Matrix Sciences) against a concatenated target-decoy database (Uniprot ver 2011_12) consisting of murine proteins and common laboratory contaminants such as trypsin. The data was searched with tryptic specificity, allowing three mis-cleavages, variable modifications of cysteine carbamidomethylation (+57.0215 daltons (Da)), methionine oxidation (+15.995 Da), lysine ubiquitylation (+114.0429 Da), 20 ppm precursor ion mass, and 0.8 Da fragment ion mass tolerance specified. Peptide spectral matches were filtered using a linear discriminant algorithm to an estimated false discovery rate (FDR) of 1%.

To identify potential DUBA substrates, ubiquitin-modified Lys-ε-Gly-Gly (K-ε-GG) peptides from digested *Duba*^{+/+} and *Duba*^{-/-} T-cell lysates were affinity purified and analysed by mass spectrometry^{25,33}. In brief, primary T cells were lysed with buffer containing 8 M urea, 20 mM HEPES, pH 8.0, 1 mM sodium orthovanadate, 2.5 mM sodium pyrophosphate and 1 mM β-glycerophosphate. Twenty milligrams of each sample (concentration based on Bradford assay) were reduced at 37 °C for 1 h in 4.5 mM DTT and alkylated with 10 mM iodoacetamide for 15 min at room temperature in the dark. Samples were diluted four times with 20 mM HEPES, pH 8.0, and digested with 10 μg ml⁻¹ trypsin overnight at 37 °C. After digestion, peptides were acidified and desalted using a Sep-Pak C18 cartridge (Waters Corp). Desalted peptides were then lyophilized for 48 h. Dried peptides were resuspended in 1.4 ml IAP buffer (Cell Signaling Technology) and incubated with pre-coupled anti-K-GG antibody beads for 2 h at 4 °C. Beads were washed twice with IAP buffer and four times with water. Peptides were eluted off antibody resin twice in 0.15% trifluoroacetic acid for 10 min at room temperature. Immunoaffinity enriched peptides were desalted using C18 STAGE-Tips³⁴. The desalted samples were subjected to mass spectrometric analysis using similar conditions as described above, but with an extended gradient and total analysis time of 120 min. Data were acquired for duplicate injections of each sample.

Tandem mass spectral results were searched using Mascot with the same search parameters as described above with the exception of semi-trypsin specificity and 25 ppm precursor ion tolerance. Peptide spectral matches were filtered using the linear discriminant algorithm to a FDR of 5% at the peptide level then to an FDR of 2% at the protein level. Modification site localization scores were generated for each K-ε-GG peptide spectra match (PSM) using a modified version of the AScore algorithm³⁵. Confidently identified peptides with ambiguous localization (modification identified on carboxy-terminal lysine of peptide by Mascot) bearing a single internal lysine residue were reported with the modification localized to the internal lysine. Peptides where the modification has been assigned to the C-terminal lysine by Mascot, and where internal lysine is absent, were discarded based on evidence suggesting trypsin does not cleave at modified lysines.

Quantification of the peptides identified with K-ε-GG modification(s) was performed using XQuant, a modified version of VistaGrande^{36,37} for processing of unlabelled samples. XQuant uses direct PSMs, precursor accurate mass, and retention time matching for quantification calculations. XQuant results were filtered to a VQ Confidence Score of 83 or greater and exported for further processing and graphical analysis using mixed-effect modelling. A mixed-effect model was fit to the area under the curve for each protein. Fold change and *P* values of mean area under the curve from knockout versus wild type were used in preparing plots.

OVA and CFA immunization. Mice were injected subcutaneously at the base of the tail with 100 µl of emulsion consisting of 100 µg OVA in 50 µl PBS and 50 µl CFA (2 mg ml⁻¹ *Mycobacterium tuberculosis* H37RA). Six days later mice were euthanized, and draining lymph nodes and spleen were collected for flow cytometric analysis. A sample size of four mice per genotype was used in each independent experiment. Female mice were used. No animals were excluded from the study. No blinding or randomization method was used in grouping the mice.

Anti-CD3 injection. Mice were injected intraperitoneally with 30 µg anti-CD3 (clone 145-2C11) in 500 µl PBS or PBS alone. After 48 h, mice were bled by retro-orbital bleed under anaesthesia and then euthanized. Small intestine oedema and inflammation were characterized by visual inspection. Sample sizes of four mice per group were used for analyses of T cells and cytokine production, group sizes of eight mice per group were used for clinical and histological analyses. No animals were excluded from any study. No blinding or randomization method was used in grouping the mice. Clinical scores were assigned as: (0) healthy small intestine; (1) <25% of small intestine inflamed; (2) <50% of small intestine inflamed; (3) <75% of small intestine inflamed; or (4) entire small intestine inflamed.

Isolation of intraepithelial lymphocytes. Peyer's patches were removed from the small intestine, which was cut open longitudinally, briefly washed with ice-cold PBS and cut into 1.5 cm pieces. Tissue was incubated in 30 ml of HBSS (5% FBS, 10 mM HEPES and 1 mM DTT) on a shaker at 250 rpm, 37 °C, for 30 min, extensively vortexed and filtered through a metal mesh. The flow-through cell suspension, which constitutes the epithelial cell content and intraepithelial lymphocytes, was centrifuged at 230g for 5 min. The cell pellet was re-suspend in 40% Percoll solution (GE Healthcare) and placed on top of a 70% Percoll solution. The gradient was centrifuged at 800g for 25 min, and intraepithelial lymphocytes at the interface were collected.

Isolation of ILCs from small intestine. Peyer's patches were removed from the small intestine, which was cut open longitudinally, briefly washed with ice-cold PBS and cut into 1.5 cm pieces. Tissues were incubated in HBSS containing 2% FBS, 10 mM HEPES, pH 7.2, 1 mM EDTA and 2 mM DTT with shaking at 230 rpm for 40 min at 37 °C. After vortex for 10 s, the tissue pieces were washed with HBSS containing 2% FBS and 10 mM HEPES, pH 7.2, minced into 1–2 mm, and incubated in HBSS containing 2% FBS, 10 mM HEPES, pH 7.2, 0.1 mg ml⁻¹ Liberase TL (Roche), and 0.15 mg ml⁻¹ DNase I (Roche) with rotating for 30 min at 37 °C. After terminating the enzyme reaction by adding DMEM containing 10% FBS, the cell suspensions were filtrated, and the cells were washed several times with DMEM. Cells were stimulated with PMA (10 ng ml⁻¹) and ionomycin (500 ng ml⁻¹) for 4 h in the presence of brefeldin A, and analysed by intracellular staining.

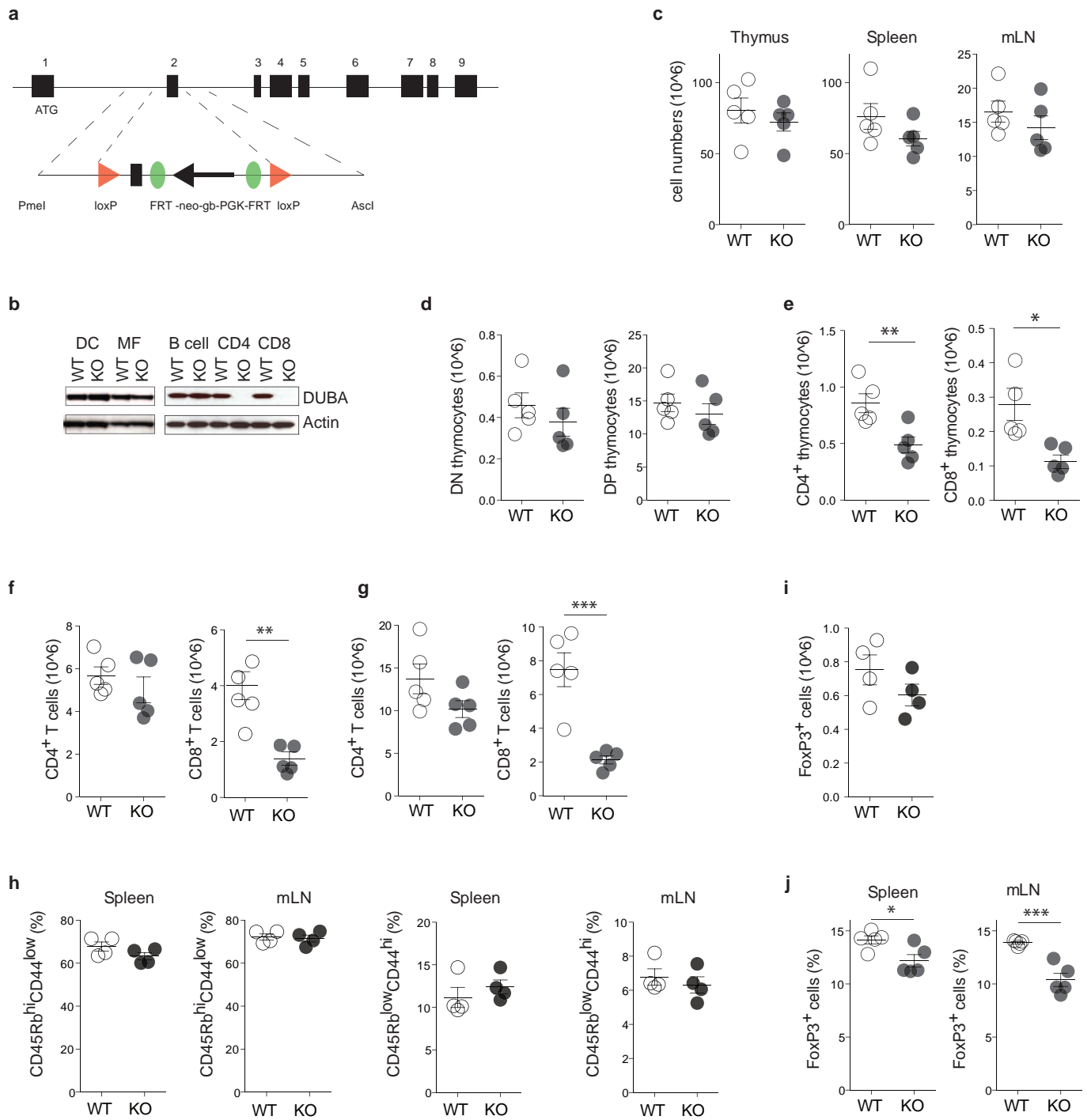
T-cell transfer colitis model. Purified naive CD4⁺ CD45Rb^{hi} CD44^{low} CD25⁻ T cells from naive C57BL/6 mice were injected intraperitoneally into *Rag2*^{-/-} recipients (400,000 cells per mouse in 200 µl sterile PBS). In some experiments, *Duba*^{+/+} and *Duba*^{-/-} T_{reg} cells were co-injected (100,000 cells per injection). Recipients were weighed weekly for 10 weeks. Animals were euthanized and the colon flushed with

ice-cold PBS. No animals were excluded from any study. Group sizes of five (PBS) and ten (all treatment groups) mice were used. No blinding or randomization method was used in grouping the mice. After manual scoring, tissues were fixed in 10% formalin. The extent of colon oedema and inflammation was determined as: (0) healthy colon; (1) <25% of colon inflamed; (2) <50% of colon inflamed; (3) <75% of colon inflamed; or (4) entire colon inflamed.

Histological lesions in the T-cell transfer colitis model included infiltration of the colonic lamina propria by T lymphocytes, with fewer macrophages, multinucleated giant cells, and neutrophils. Inflammatory infiltrates separated crypts and rarely extended into the tunica muscularis. Crypt hyperplasia and, less frequently, mucosal erosions and focal necrosis were present in areas of more severe inflammation. Histological lesions were scored blindly according to the following criteria: (0) normal colon, (1) minimal inflammation with minimal to no separation of crypts (generally focal affecting <10% of mucosa), (2) mild inflammation with mild separation of crypts (generally affecting 11–25% of mucosa or mild, diffuse inflammatory infiltrates were minimal separation of crypts), (3) moderate inflammation with separation of crypts ± focal effacement of crypts (generally affecting 26–50% of mucosa or diffuse, moderate separation of crypts), (4) extensive inflammation with marked separation and effacement of crypts (generally affecting 51–75% of mucosa), and (5) diffuse inflammation with marked separation and effacement of crypts (generally affecting >76% of mucosa).

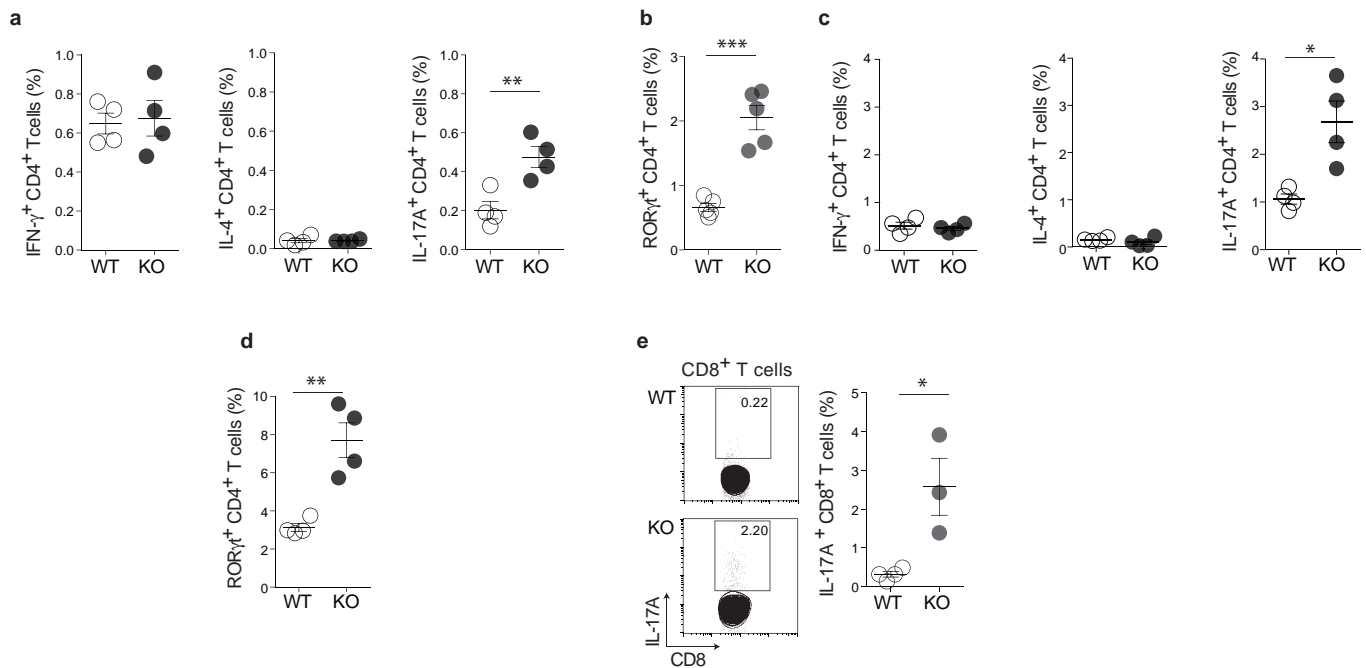
Statistical analysis. Statistical significance was determined with an unpaired *t*-test using the GraphPad Prism software. No statistical method was used to predetermine sample size.

30. siRNA stabilization prolongs gene knockdown in primary T lymphocytes. *Eur. J. Immunol.* **38**, 2616–2625 (2008).
31. Collison, L. W. & Vignali, D. A. A. *In vitro* Treg suppression assays. *Methods Mol. Biol.* **707**, 21–37 (2011).
32. Zheng, Y. *et al.* Interleukin-22, a T_H17 cytokine, mediates IL-23-induced dermal inflammation and acanthosis. *Nature* **445**, 648–651 (2007).
33. Anania, V. G. *et al.* Peptide level immunoaffinity enrichment enhances ubiquitination site identification on individual proteins. *Mol. Cell. Proteomics* **13**, 145–156 (2014).
34. Rappsilber, J., Ishihama, Y. & Mann, M. Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal. Chem.* **75**, 663–670 (2003).
35. Beausoleil, S. A., Villén, J., Gerber, S. A., Rush, J. & Gygi, S. P. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nature Biotechnol.* **24**, 1285–1292 (2006).
36. Bakalarski, C. E. *et al.* The impact of peptide abundance and dynamic range on stable-isotope-based quantitative proteomic analyses. *J. Proteome Res.* **7**, 4756–4765 (2008).
37. Kirkpatrick, D. S. *et al.* Phosphoproteomic characterization of DNA damage response in melanoma cells following MEK/PI3K dual inhibition. *Proc. Natl Acad. Sci. USA* **110**, 19426–19431 (2013).



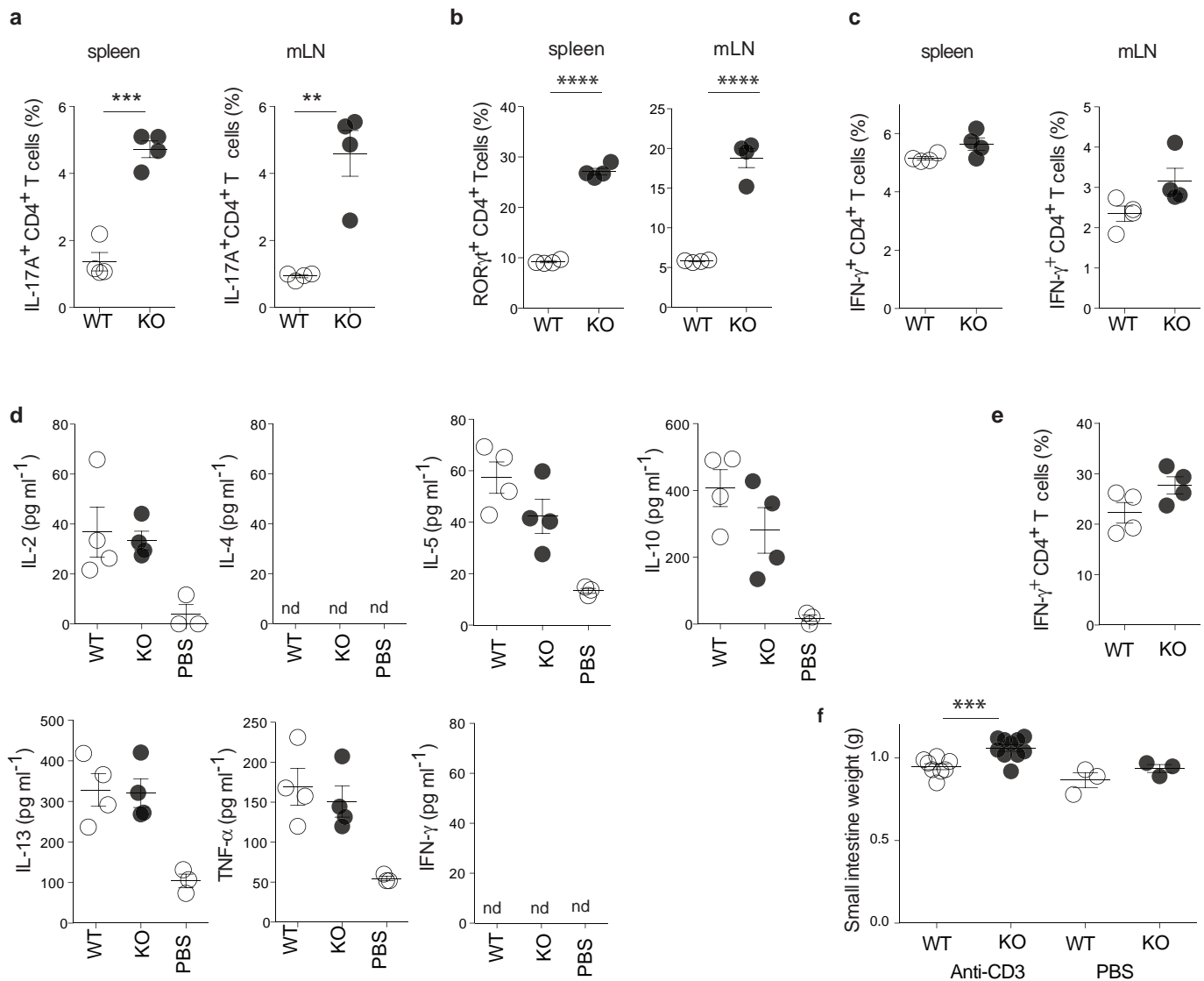
Extended Data Figure 1 | Analysis of the T-cell compartment in *Duba*^{fl/fl} CD4-Cre mice aged 6 weeks. **a**, Targeting strategy for the *Duba*^{fl} allele. The PGK-neo selection cassette was removed with an Flp deleter strain. **b**, Immunoblots of immune cell subsets from *Duba*^{+/+} CD4-Cre (wild-type, WT) and *Duba*^{fl/fl} CD4-Cre (knockout, KO) mice. CD11c⁺ dendritic cells (DC) and peritoneal macrophages (MF) were stimulated with 50 ng ml⁻¹ lipopolysaccharide (LPS) overnight. **c–j**, Knockout and wild-type mice were aged 6 weeks. *n* = 5 mice per genotype. Leukocyte numbers in the thymus,

spleen and mesenteric lymph node (**c**), CD4[−]CD8[−] (DN) and CD4⁺CD8⁺ (DP) cells in the thymus enumerated by cell counting and flow cytometry (**d**), Numbers of CD4⁺CD8[−] and CD4[−]CD8⁺ cells in thymus (**e**), in mesenteric lymph nodes (mLN) (**f**) and spleen (**g**). **h**, Percentages of naive (CD45Rb^{hi}CD44^{low}) and memory (CD45Rb^{low}CD44^{hi}) CD4⁺ T cells. **i**, FOXP3⁺ T cells in the thymus. **j**, Percentage of FOXP3⁺ T cells in the spleen and mesenteric lymph nodes. Each circle represents one mouse. Error bars, s.e.m. **P* < 0.05, ***P* < 0.01, ****P* < 0.001 (unpaired Student's *t*-test).



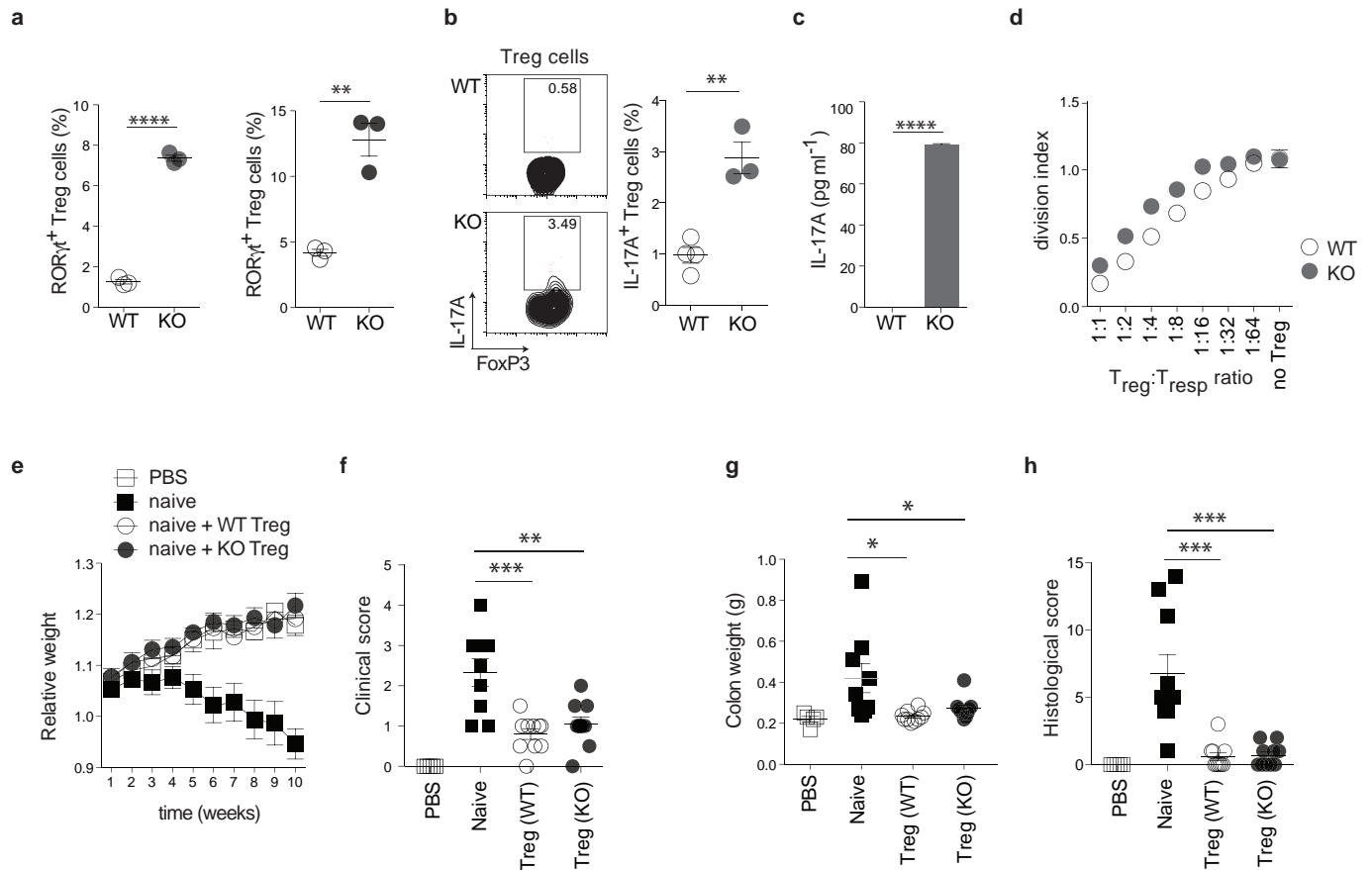
Extended Data Figure 2 | Increased IL-17A production by CD4⁺ and CD8⁺ T cells from *Duba*^{n/n} CD4-Cre mice. **a, b**, Percentage of splenic CD4⁺ T cells from wild-type (*Duba*^{+/+} CD4-Cre) or knockout (*Duba*^{n/n} CD4-Cre) mice that stained positive for intracellular IFN- γ , IL-4 or IL-17A (**a**) or ROR γ t (**b**) by flow cytometry. **c, d**, Percentage of CD4⁺ T cells expressing IFN- γ , IL-4 and IL-17A (**c**) or ROR γ t (**d**) after collection on day 8 after immunization with

OVA/CFA, and subsequent culture with PMA and ionomycin for 4 h. **e**, Percentage of splenic CD8⁺ T cells from knockout and wild-type mice that stained positive for intracellular IL-17A by flow cytometry. Representative contour plots are shown. Each circle represents one mouse. Error bars, s.e.m. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$ (unpaired Student's *t*-test).



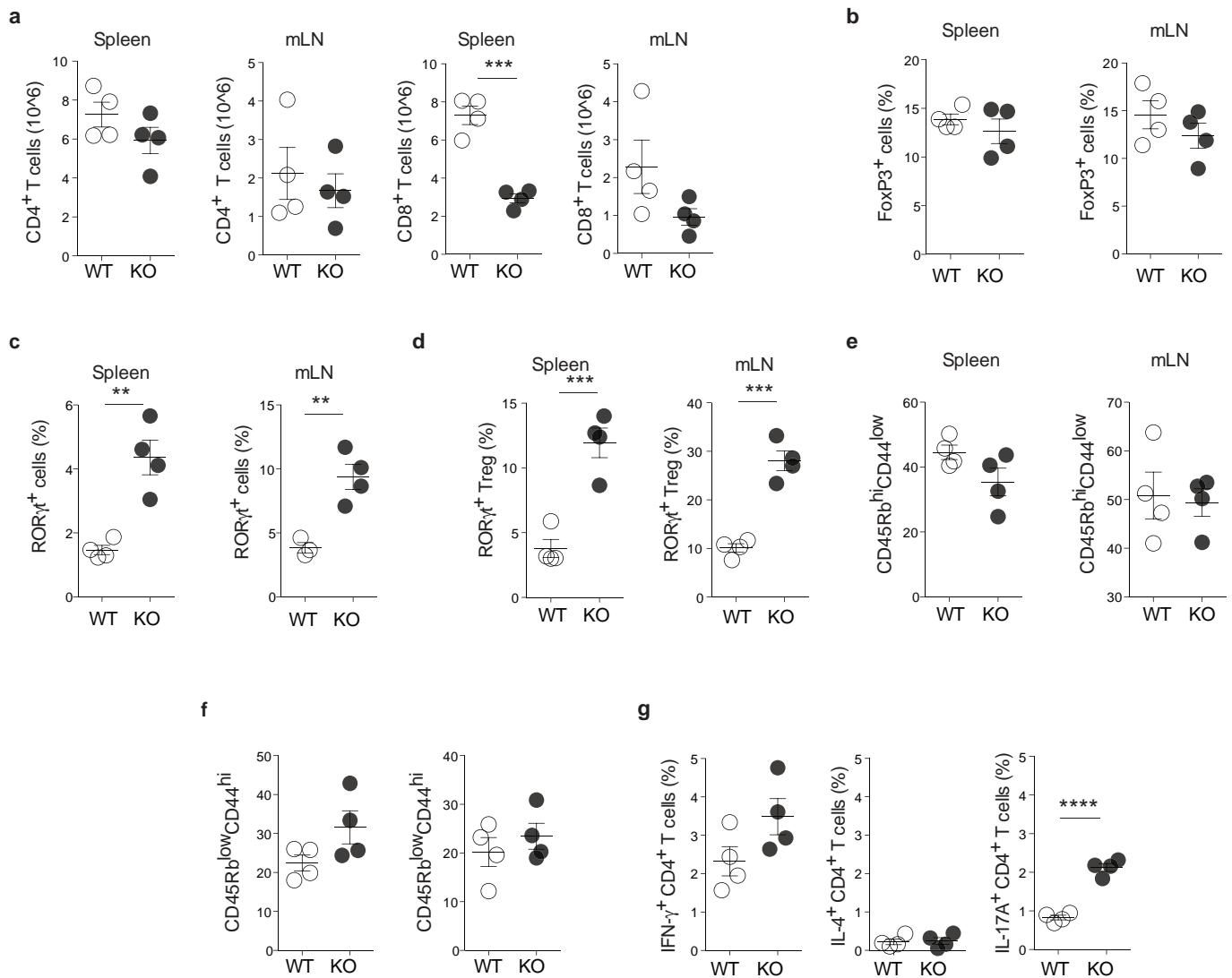
Extended Data Figure 3 | Cytokine production by *Duba*^{+/+} CD4-Cre and *Duba*^{fl/fl} CD4-Cre mice injected with anti-CD3 antibodies. **a–c**, CD4⁺ T cells expressing IL-17A (**a**), RORγt (**b**) or IFN-γ (**c**) by intracellular staining and flow cytometry. Cells from the spleen or mesenteric lymph node were isolated from wild-type (*Duba*^{+/+} CD4-Cre) or knockout (*Duba*^{fl/fl} CD4-Cre) mice at 48 h after injection with anti-CD3 antibodies and then stimulated with PMA and ionomycin for 4 h. Representative contour plots are shown. **d**, Serum cytokines at 48 h after injection of anti-CD3 antibodies. PBS indicates

wild-type mice injected with vehicle alone. nd, not detected. **e**, Intraepithelial lymphocytes from the small intestine producing IFN-γ based on intracellular staining and flow cytometry. Cells were collected at 48 h after injection and stimulated with PMA and ionomycin for 4 h. **f**, Small intestine weights at 48 h after injection of anti-CD3 antibodies. Each circle represents one mouse. Error bars, s.e.m. **P* < 0.05, ***P* < 0.01, ****P* < 0.001 (unpaired Student's *t*-test). Data are representative of 2 (**a–d, f**) or 3 (**e**) independent experiments.



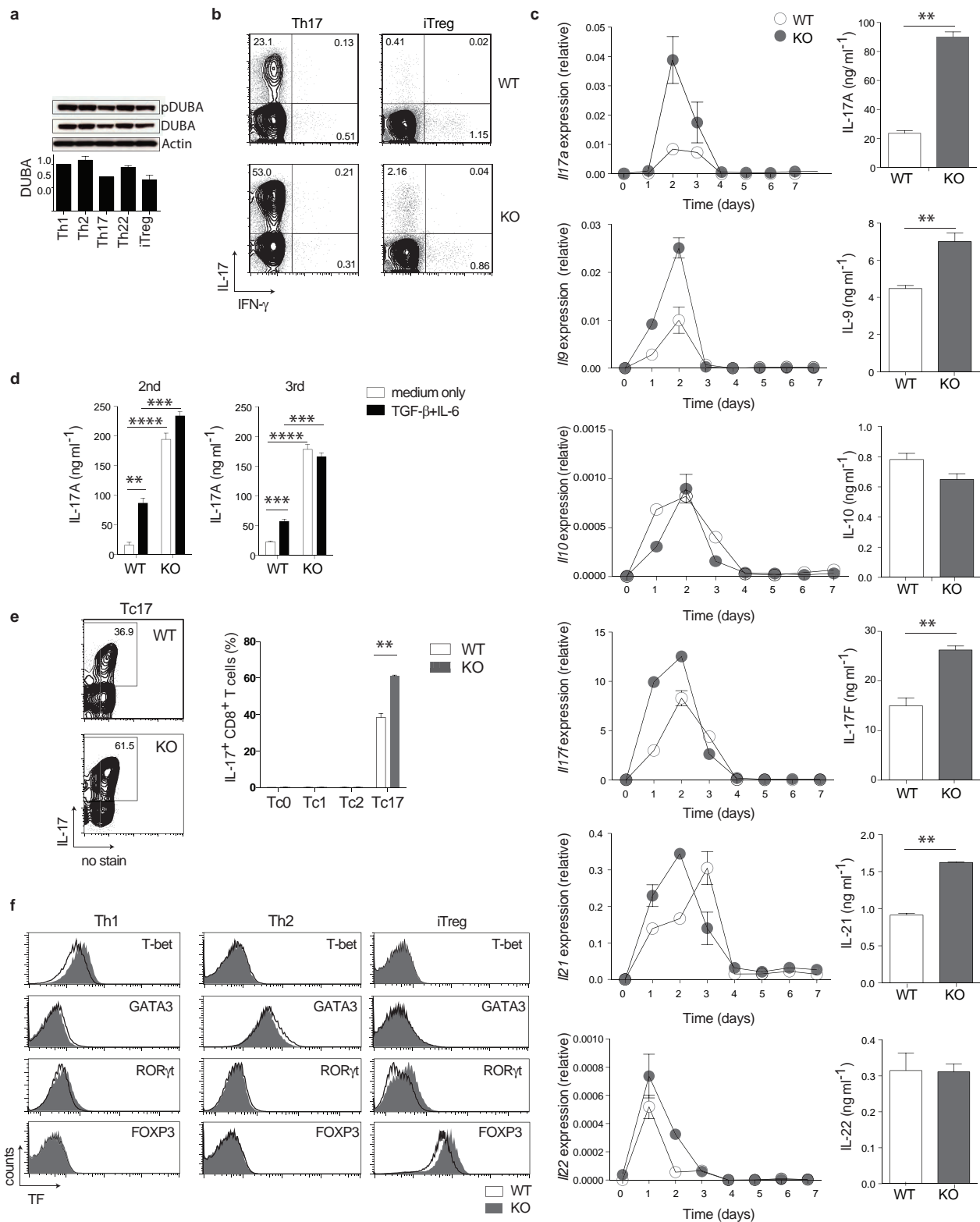
Extended Data Figure 4 | DUBA-deficient T_{reg} cells express RORγt and produce IL-17A, but are suppressive *in vitro* and *in vivo*. **a, b**, Percentage of RORγt⁺ (a) or IL-17A-producing (b) T_{reg} cells in the spleen or mesenteric lymph node of naive mice. **c**, IL-17A secretion by CD25⁺ T_{reg} cells after stimulation with anti-CD3 and anti-CD28 antibodies for 24 h. Data are representative of two independent experiments. **d**, Proliferation of C57BL/6 CD4⁺ T cells (responder T cells; T_{resp}) co-cultured with T_{reg} cells in the ratios indicated. Division index denotes the average number of cell divisions that a

cell in the original population has undergone. Data are representative of two independent experiments. **e**, Weight relative to time zero of Rag2^{-/-} mice after transfer of either naive CD4⁺ T cells alone or together with DUBA^{+/+} or DUBA^{-/-} T_{reg} cells. **f**, Animal clinical scores. **g**, Colon weights at 10 weeks after T-cell transfer. **h**, Colon histology scores at 10 weeks. *n* = 10 (cell transfer groups), *n* = 5 (PBS). Error bars, s.e.m. **P* < 0.05, ***P* < 0.01, ****P* < 0.001 (unpaired Student's *t*-test). Data are representative of two independent experiments.



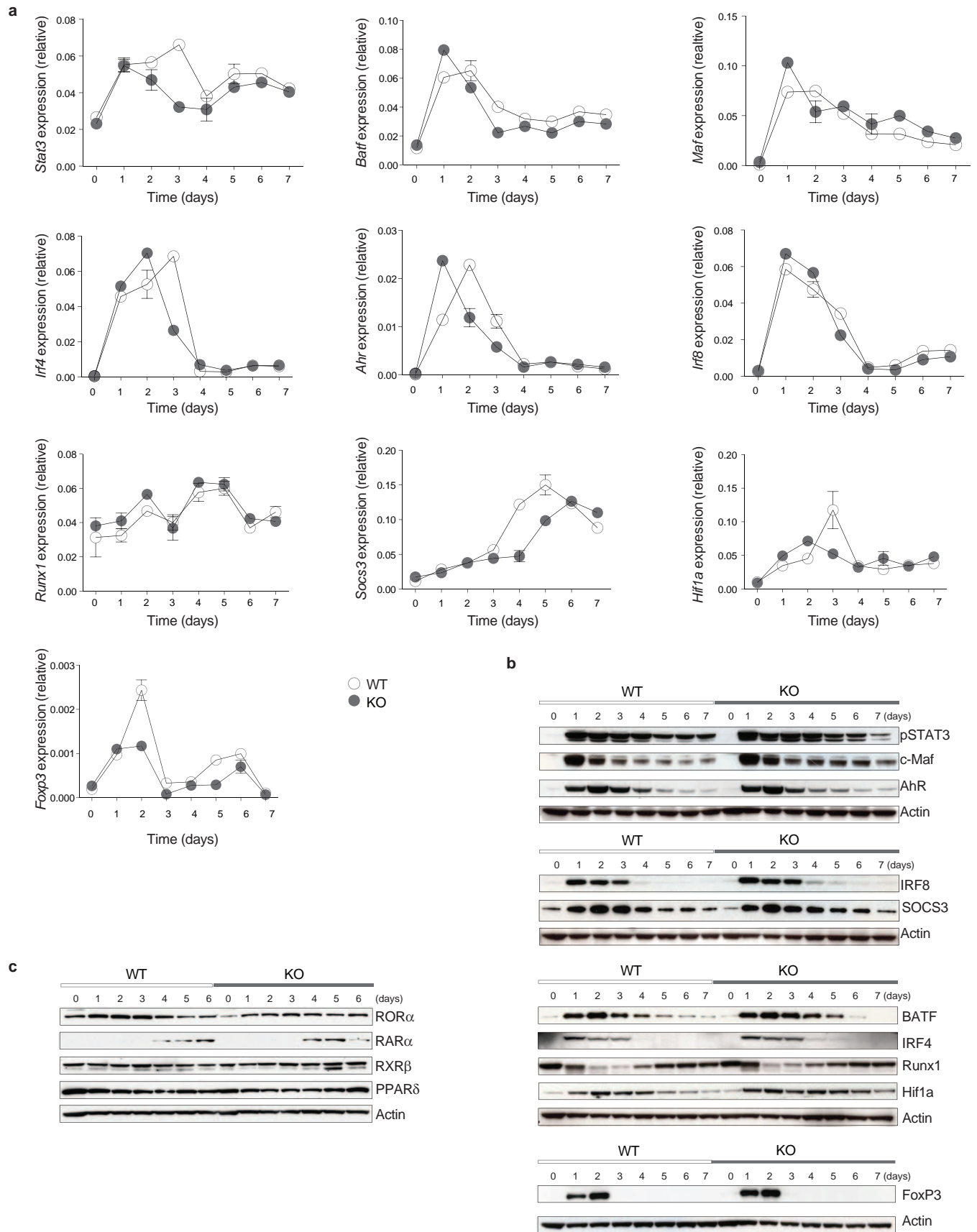
Extended Data Figure 5 | Composition of the T-cell compartment in *Duba^{nl/n} CD4-Cre* mice aged one year. **a**, $CD4^{+}$ and $CD8^{+}$ T cells enumerated by cell counting and flow cytometry. **b–f**, Percentages of FOXP3⁺ T_{reg} cells (**b**), $ROR\gamma^{+}$ T cells (**c**), $ROR\gamma^{+}$ T_{reg} cells (**d**), naive ($CD45Rb^{hi}CD44^{low}$) $CD4^{+}$ T cells (**e**), and memory ($CD45Rb^{low}CD44^{hi}$) $CD4^{+}$ T cells (**f**) in the

spleen and mesenteric lymph node. **g**, Percentages of IFN- γ , IL-4 or IL-17-producing $CD4^{+}$ T cells in spleen. Each circle represents one mouse. Error bars, s.e.m. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, **** $P < 0.0001$ (unpaired Student's *t*-test).



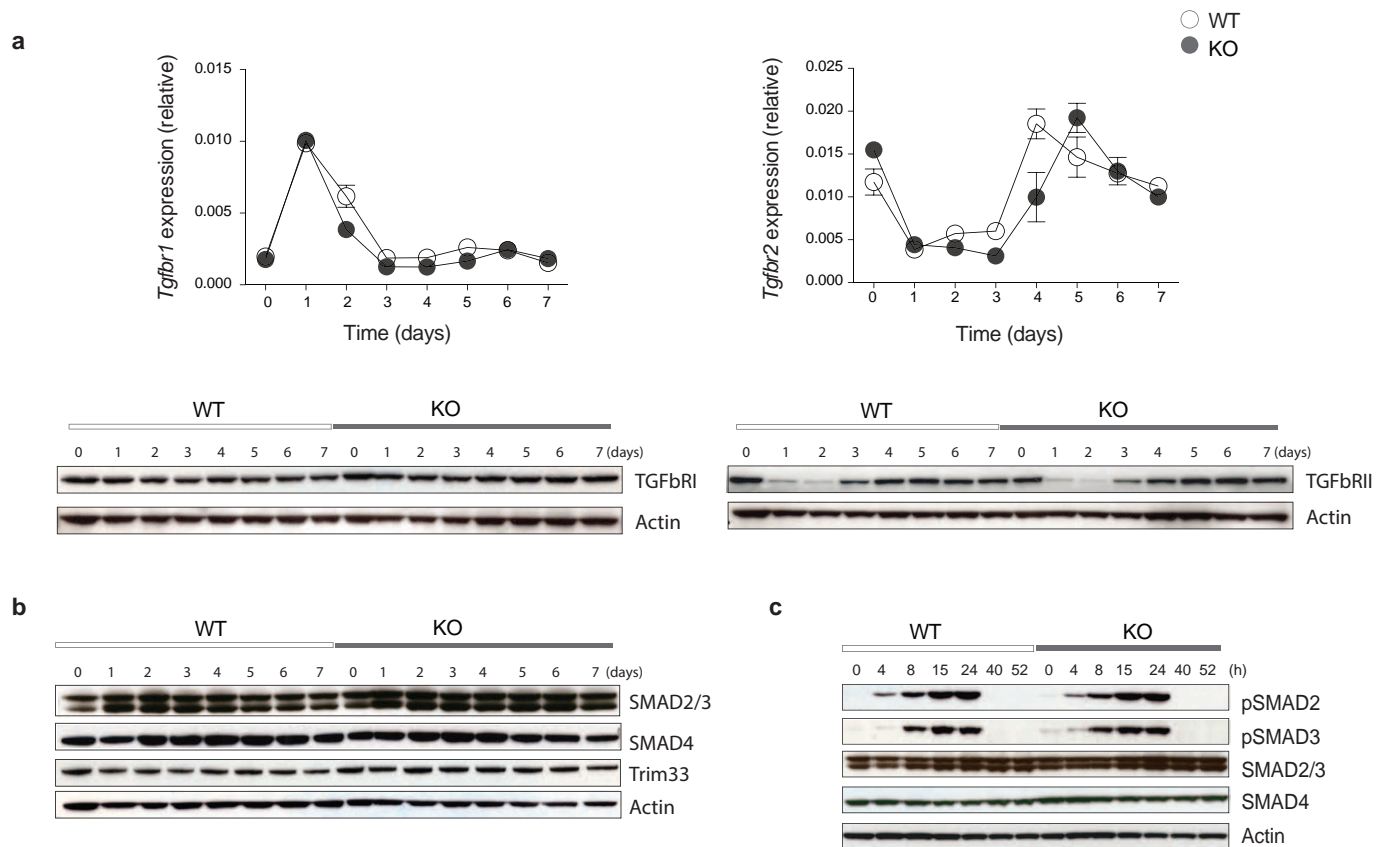
Extended Data Figure 6 | Cytokine production and lineage transcription factor expression in DUBA-deficient T cells. **a**, Immunoblots of T-helper subsets after 24 h of culture. Graphs indicate relative DUBA abundance from densitometry measurements. Error bars, s.e.m. of three independent experiments. **b**, Representative contour plots indicating the percentage of IL-17A-expressing T_H17 and iT_{reg} cells. **c**, Expression of *Il17a*, *Il21*, *Il9*, *Il22* and *Il10* mRNAs by *Duba*^{+/-} and *Duba*^{-/-} T cells that were stimulated with anti-CD3 and anti-CD28 antibodies in the presence of TGF- β and IL-6. Secreted cytokines were measured on the third day of culture. **d**, IL-17A secretion by

Duba^{+/-} or *Duba*^{-/-} naive CD4⁺ T cells after culture with TGF- β and IL-6 for 6 days and then secondary stimulation with anti-CD3 and anti-CD28 antibodies for 2 days. After a 6-day secondary stimulation, the cells were re-stimulated again. **e** Percentage of CD8⁺ T cells expressing IL-17A after 5 days of culture under T_H0 , T_H1 , T_H2 and T_H17 polarizing conditions by flow cytometry. Representative contour plots under T_H17 conditions are shown. **f**, Flow cytometric analysis of T-bet, GATA3, ROR γ t and FOXP3 expression in T_H1 , T_H2 and iT_{reg} cells. Error bars, s.e.m. of triplicate measurements. Data are representative of two independent experiments.



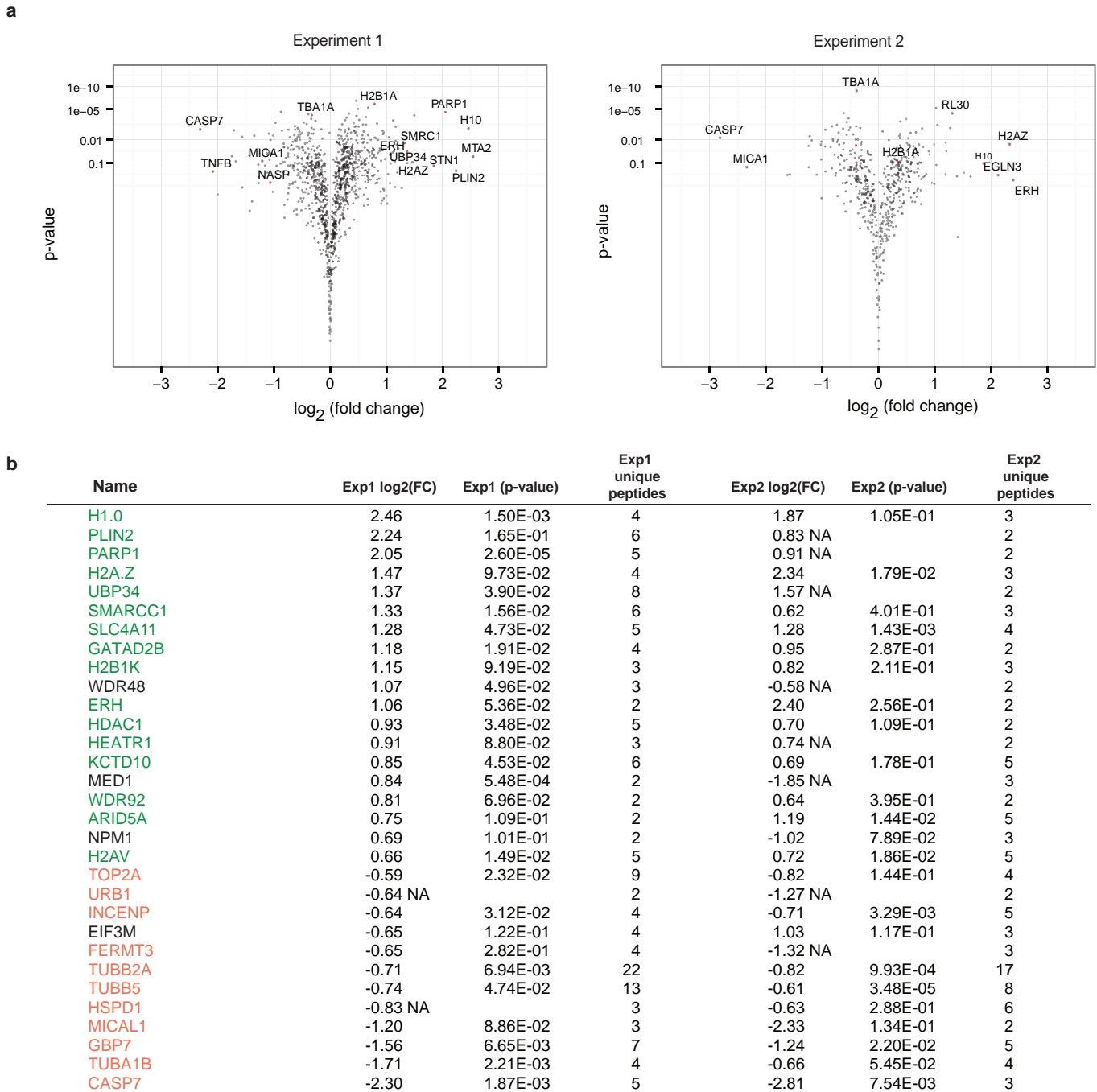
Extended Data Figure 7 | Normal expression of T_H17 factors in DUBA-deficient T_H17 cells. **a–c**, Expression of T_H17 factors by $Duba^{+/+}$ and $Duba^{-/-}$ $CD4^+$ T cells that were stimulated with anti-CD3 and anti-CD28 antibodies in the presence of TGF- β and IL-6. Graphs show mRNA expression

(a) and immunoblots indicate protein expression (b), immunoblots indication expression of retinoid acid receptors (c). Data are representative of two independent experiments.



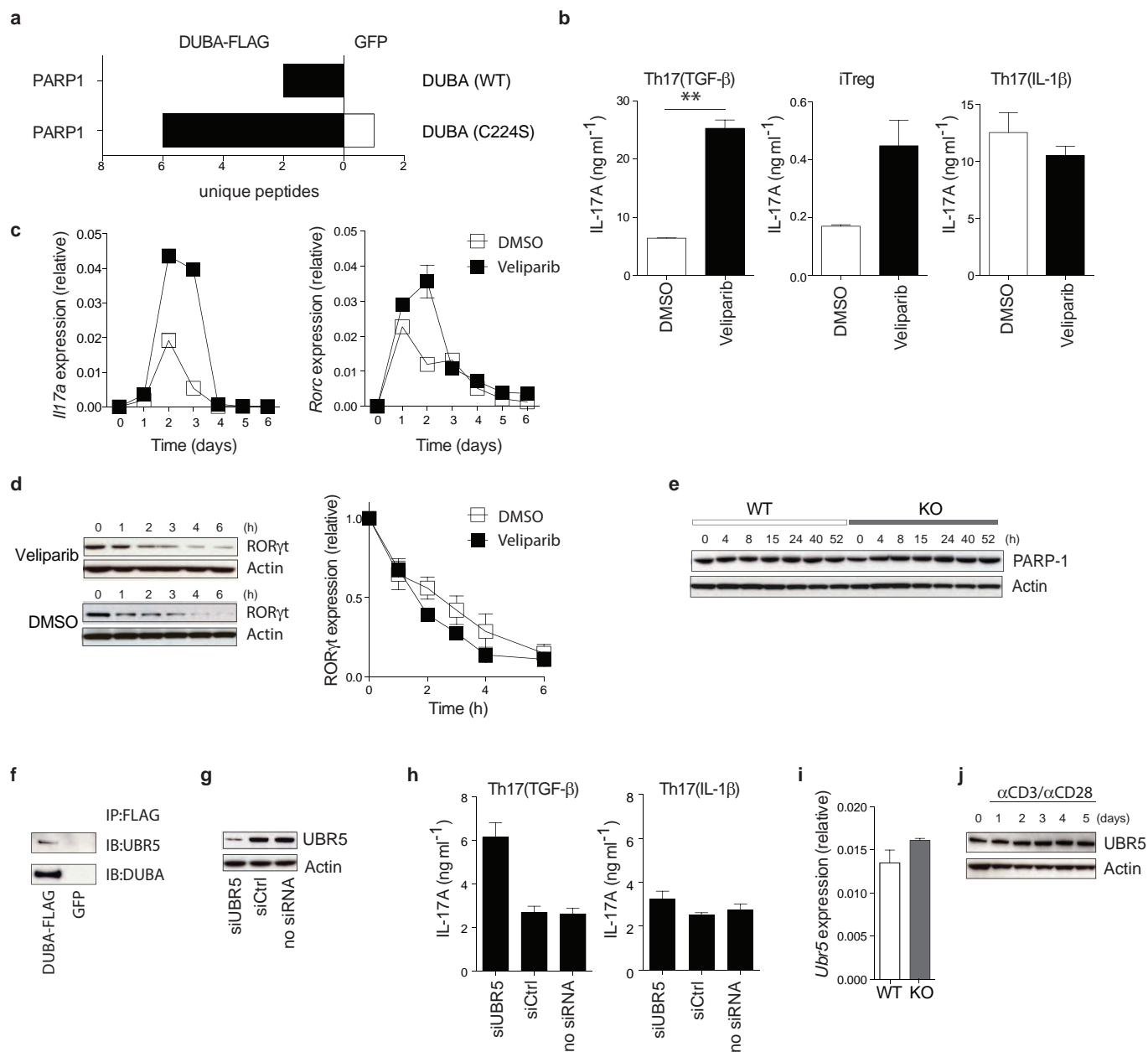
Extended Data Figure 8 | Normal TGF- β receptor expression and SMAD activation in DUBA-deficient TH17 cells. **a**, Expression of TGF- β RI and TGF- β RII in *Duba*^{+/+} and *Duba*^{-/-} TH17 cells after culture for the times indicated. Graphs indicate mRNA expression, whereas immunoblots indicate

protein expression. Error bars, s.e.m. of duplicate measurements. Data are representative of two independent experiments. **b**, Immunoblots of the cells in **a**. **c**, Immunoblots of *Duba*^{+/+} or *Duba*^{-/-} TH17 cells after stimulation with TGF- β and IL-6.



Extended Data Figure 9 | Identification of DUBA substrates in TH17 cells by K-ε-GG peptide analysis. **a**, Volcano plots of ubiquitinated proteins identified after immunoprecipitating K-ε-GG peptides from *Duba*^{+/+} and *Duba*^{-/-} TH17 cells after 4 days of culture. *P* values are plotted against

fold-change in number of peptides for each protein. **b**, Proteins >1.5-fold differentially ubiquitinated in *Duba*^{+/+} and *Duba*^{-/-} TH17 cells (green and red denote increased and decreased ubiquitylation in the knockout, respectively).



Extended Data Figure 10 | The roles of PARP-1 and UBR5 in the regulation of IL-17A production in T cells. **a**, Mass spectrometry detection of PARP-1 co-immunoprecipitated with Flag-tagged wild-type DUBA or a DUBA(Cys224Ser) mutant expressed in TH17 cells. Graph indicates the number of unique peptides in experimental samples (black) or control cells expressing GFP (white). **b**, IL-17A production by TH17 cells or iTreg cells (TGF- β plus IL-2) after 3 days of culture in DMSO vehicle or veliparib. **c**, Expression of *Il17a* and *Rorc* mRNA in TH17 cells treated with veliparib or DMSO vehicle. **d**, Immunoblots of ROR γ t in TH17 cells cultured for 2 days with either DMSO or veliparib, and then treated for the indicated times in cycloheximide. Graphs indicate relative ROR γ t abundance from densitometry measurements. Error bars, s.e.m. of three independent experiments.

e, Immunoblot of *Duba*^{+/+} and *Duba*^{-/-} TH17 cells cultured for the times indicated. Error bars, s.e.m. of triplicate measurements. ****** $P < 0.01$ (unpaired Student's *t*-test). **f**, Immunoblot of UBR5 co-immunoprecipitated with Flag-tagged DUBA from retrovirally transduced CD4⁺ T cells. Control cells were transduced with GFP-expressing virus. **g**, Immunoblot of UBR5 in wild-type TH17 cells transfected with the siRNA indicated. **h**, IL-17A production by wild-type TH17 cells transfected with the siRNA indicated. Error bars, s.e.m. of two independent experiments. **i**, Expression of *Ubr5* mRNA in CD4⁺ T cells. Error bars, s.e.m. of three independent experiments. **j**, Immunoblots of wild-type T cells stimulated with anti-CD3 or anti-CD28 antibodies for the times indicated.

Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution

Peter Eirew^{1,2*}, Adi Steif^{1,2*}, Jaswinder Khattra^{1,2*}, Gavin Ha^{1,2}, Damian Yap^{1,2}, Hossein Farahani^{1,2}, Karen Gelmon³, Stephen Chia³, Colin Mar³, Adrian Wan¹, Emma Laks^{1,2}, Justina Biele^{1,2}, Karey Shumansky¹, Jamie Rosner¹, Andrew McPherson^{1,2}, Cydney Nielsen^{1,2}, Andrew J. L. Roth^{1,2}, Calvin Lefebvre^{1,2}, Ali Bashashati^{1,2}, Camila de Souza¹, Celia Siu¹, Radhouane Aniba^{1,2}, Jazmine Brimhall¹, Arusha Oloumi^{1,2}, Tomo Osako^{1,2}, Alejandra Bruna^{4,5}, Jose L. Sandoval^{4,5}, Teresa Algras^{1,2}, Wendy Greenwood^{4,5}, Kaston Leung^{6,7}, Hongwei Cheng^{8,9}, Hui Xue^{8,9}, Yuzhuo Wang^{8,9}, Dong Lin^{8,9}, Andrew J. Mungall¹⁰, Richard Moore¹⁰, Yongjun Zhao¹⁰, Julie Lorette¹¹, Long Nguyen^{12,13}, David Huntsman^{2,11}, Connie J. Eaves^{12,13}, Carl Hansen^{6,7}, Marco A. Marra¹⁰, Carlos Caldas^{4,5}, Sohrab P. Shah^{1,2,10} & Samuel Aparicio^{1,2,10,11}

Human cancers, including breast cancers, comprise clones differing in mutation content. Clones evolve dynamically in space and time following principles of Darwinian evolution^{1,2}, underpinning important emergent features such as drug resistance and metastasis^{3–7}. Human breast cancer xenograftment is used as a means of capturing and studying tumour biology, and breast tumour xenografts are generally assumed to be reasonable models of the originating tumours^{8–10}. However, the consequences and reproducibility of engraftment and propagation on the genomic clonal architecture of tumours have not been systematically examined at single-cell resolution. Here we show, using deep-genome and single-cell sequencing methods, the clonal dynamics of initial engraftment and subsequent serial propagation of primary and metastatic human breast cancers in immunodeficient mice. In all 15 cases examined, clonal selection on engraftment was observed in both primary and metastatic breast tumours, varying in degree from extreme selective engraftment of minor (<5% of starting population) clones to moderate, polyclonal engraftment. Furthermore, ongoing clonal dynamics during serial passaging is a feature of tumours experiencing modest initial selection. Through single-cell sequencing, we show that major mutation clusters estimated from tumour population sequencing relate predictably to the most abundant clonal genotypes, even in clonally complex and rapidly evolving cases. Finally, we show that similar clonal expansion patterns can emerge in independent grafts of the same starting tumour population, indicating that genomic aberrations can be reproducible determinants of evolutionary trajectories. Our results show that measurement of genomically defined clonal population dynamics will be highly informative for functional studies using patient-derived breast cancer xenograftment.

To evaluate xenograft clonal dynamics (see Supplementary Table 1 for definitions of terms used) we generated 30 xenograft lines by serially transplanting (up to 16 generations over 3 years) breast cancer tissue organoid suspensions from 55 patients (Extended Data Fig. 1, Supplementary Table 2 and Supplementary Fig. 1) into highly immunodeficient *NOD/SCID/Il2rg^{-/-}* (NSG) and *NOD/Rag1^{-/-}Il2rg^{-/-}* (NRG) mice¹¹ (details in the Supplementary Information). We carried out massively parallel whole-genome shotgun sequencing (WGSS) on DNA from xenograft passages of 15 patient lines (10 primary tumour-derived and five pleural effusion-derived), along with matched patient tumour and

normal DNA (47 samples total, median sequencing depth 45.1, Supplementary Table 3). For these, plus 56 additional xenograft passage samples, we validated 3,187 somatic single nucleotide variant (SNV) positions (100–300 per tumour-xenograft series) and 132 structural variant positions by targeted-amplicon deep sequencing (Supplementary Tables 4–6), quantifying allele ratios to a high level of precision. We surveyed the copy-number alteration (CNA) and loss of heterozygosity (LOH) landscapes using Affymetrix SNP Array 6.0 (Supplementary Tables 7 and 8). The mutation load of somatic SNVs (range: 4.3–27.7 × 10³ genome-wide; 57–1,040 in coding regions), CNA and LOH (34–67% of genome), and structural variants in the 15 tumour-xenograft series (Supplementary Figs 2 and 3 and Supplementary Table 9) were consistent with previous genome-wide breast cancer studies^{4,12–17}, although low tumour cellularity hindered mutation discovery in case numbers SA429 and SA496 originating tumours. Tumour-xenograft pairs displayed comparable nucleotide substitution patterns (Supplementary Figs 2 and 4), suggesting that mutational processes are maintained post-engraftment.

To determine the extent of evolution in the SNV landscape, we first compared the genome-wide variant allele prevalences (the proportion of aligned reads at the SNV position with the variant base, see Supplementary Table 1) from WGSS data in xenograft relative to tumour (SA429 and SA496 excluded due to low tumour cellularity). As expected, sizeable proportions (range: 53.0–92.9%) of high-confidence SNVs are shared in tumour-xenograft pairs, with prevalences lying on a scatter plot diagonal indicating neutral dynamics (Extended Data Fig. 2a and Supplementary Figs 5a and 6). Notably, all 15 samples also show clusters of SNVs prevalent in the xenograft while at or below the limit of detection in the tumour (range: 6.5–32.1% of SNVs, see for example, SA494, SA495 and SA499) and vice versa (range: 0.2–19.4%, see for example, SA494, SA495 and SA500), implying clonal selection on initial engraftment. Tumours and xenografts from SA494, SA495, SA499, SA500 and SA530 also exhibited substantial differences in structural variant content (Supplementary Figs 3 and 7).

To resolve clonal dynamics and genotypes, we applied a Bayesian clustering model (PyClone^{4,18}) to SNV variant allele prevalences measured by targeted deep sequencing, accounting for the effect of copy number, LOH status and cellularity. SNVs with co-varying estimates of cellular prevalence (the proportion of tumour or xenograft cells bearing the

¹Department of Molecular Oncology, BC Cancer Agency, 675 West 10th Avenue, Vancouver, British Columbia V5Z 1L3, Canada. ²Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, British Columbia V6T 2B5, Canada. ³Department of Medical Oncology, BC Cancer Agency, 600 West 10th Avenue, Vancouver, British Columbia V5Z 4E6, Canada. ⁴Department of Oncology, University of Cambridge, Hills Road, Cambridge CB2 2XZ, UK. ⁵Cancer Research UK Cambridge Research Institute, University of Cambridge, Li Ka Shing Centre, Cambridge CB2 0RE, UK. ⁶Centre for High-Throughput Biology, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada. ⁷Department of Physics and Astronomy, University of British Columbia, Vancouver, British Columbia V6T 1Z1, Canada. ⁸Department of Experimental Therapeutics, BC Cancer Agency, 675 West 10th Avenue, Vancouver, British Columbia V5Z 1L3, Canada. ⁹The Vancouver Prostate Centre, Vancouver General Hospital and Department of Urologic Sciences, University of British Columbia, Vancouver, British Columbia V5Z 1M9, Canada. ¹⁰Michael Smith Genome Sciences Centre, Vancouver, British Columbia V5Z 1L3, Canada. ¹¹Centre for Translational and Applied Genomics, BC Cancer Agency, 600 West 10th Avenue, Vancouver, British Columbia V5Z 4E6, Canada. ¹²Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia V6T 1Z3, Canada. ¹³Terry Fox Laboratory, BC Cancer Agency, Vancouver, British Columbia V5Z 1L3, Canada.

*These authors contributed equally to this work.

mutation) across all time points are grouped into putative mutation clusters (Supplementary Table 1). Consistent with the raw variant allele prevalence measurements, several cases contained mutation clusters with high (75–100%) prevalences in the xenografts and low (0–15%) prevalences in the tumours, implying expansion of initially minor clones to dominate the xenograft (for example, clusters 3, 4, 3, 2, 8, 2 and 2 in SA494, SA495, SA500, SA530, SA532, SA533 and SA535, respectively) (Extended Data Fig. 2b and Supplementary Fig. 5b). Other series (SA493, SA499, SA501, SA531, SA534 and SA536) demonstrated non-neutral clonal dynamics but involving alleles occupying much smaller proportions of total cellular populations. Notably, polyclonal population structure specific to the xenograft was observed after initial expansion in SA493, SA494, SA495, SA500 and SA531, suggesting that initial selection on engraftment remains permissive to additional clonal evolution (Extended Data Fig. 2b and Supplementary Fig. 5b). Polyclonal engraftment was evident in SA493, SA501, SA531 and SA532, suggesting that multiple clones maintained their fitness post-engraftment.

Analogously, we analysed clonal dynamics using CNAs as clonal marks, applying a probabilistic model (TITAN¹⁹) that infers CNA and LOH from WGSS data, accounting for mixtures of tumour and normal cells and reporting estimates of mutation cellular prevalence and mutation cluster membership (Supplementary Table 10). Despite conservation of complex disruptions, such as chromothripsis in SA429 (Supplementary Fig. 8) and breakage–fusion–bridge cycles in SA429 and SA494 (Supplementary Figs 9 and 10), we identified substantial differences in copy-number architecture between tumour and xenograft in all cases (Extended Data Fig. 2c and Supplementary Fig. 5c). These included a xenograft-specific deletion event containing *TP53* (in SA500) that coincided with retention of a somatic SNV (Supplementary Fig. 11 and Supplementary Table 6). Notably, the predominant clonal dynamic (minor subclone

expansion in SA494, SA495, SA532 and SA533; polyclonal engraftment in SA493 and SA501) mirrored those seen in SNV space.

We next asked how clonal dynamics differ after initial engraftment, using PyClone predictions over serial passage generations spanning up to 3 years (Extended Data Fig. 1). We distinguished statistically significant directional clonal dynamics by testing the overlap of 90% credible intervals derived from Bayesian posterior probability distributions (Fig. 1). Cases showing strongest clonal dynamics in the first engraftment passages (for example, SA500, SA530, SA494 and SA535) exhibited more stable prevalence over subsequent passages. In contrast, cases showing moderate initial clonal dynamics showed more marked subsequent dynamics (for example, mutation clusters 2, 3 and 8 of SA501), in some cases leading to gradual expansion of a minor clone to dominate the xenograft over serial passages. We noted examples of all oestrogen receptor/HER2 subtypes and primary/metastatic cancers evolving by these two different modes. Some mutation clusters showed non-dynamic patterns over time (for example, clusters 1, 4 and 6 of SA500, clusters 1–3, 5, 7, 9 and 10 in SA532, as well as the highest prevalence clusters representing putative ancestral mutations that remained invariant, as expected). For two cases we noted preferential engraftment of initial transplants in mammary fat pad over subrenal sites (SA496 4 of 4 mammary fat pad versus 0 of 4 subrenal; SA429 2 of 4 mammary fat pad versus 0 of 4 subrenal, Extended Data Fig. 1). However, transplant site changes in established xenografts were not associated with unusually strong clonal dynamics (Fig. 1, see SA495 X3–4, SA499 X3–4, SA429 X1–2 and SA496 X1–2, where X denotes the xenograft passage).

To validate the population-based inference of mutation clusters and clonal genotypes directly, we carried out single-cell analyses of cases SA494 (an example of extreme initial selection) and SA501 (complex post-engraftment clonal dynamics). We performed multiplexed targeted

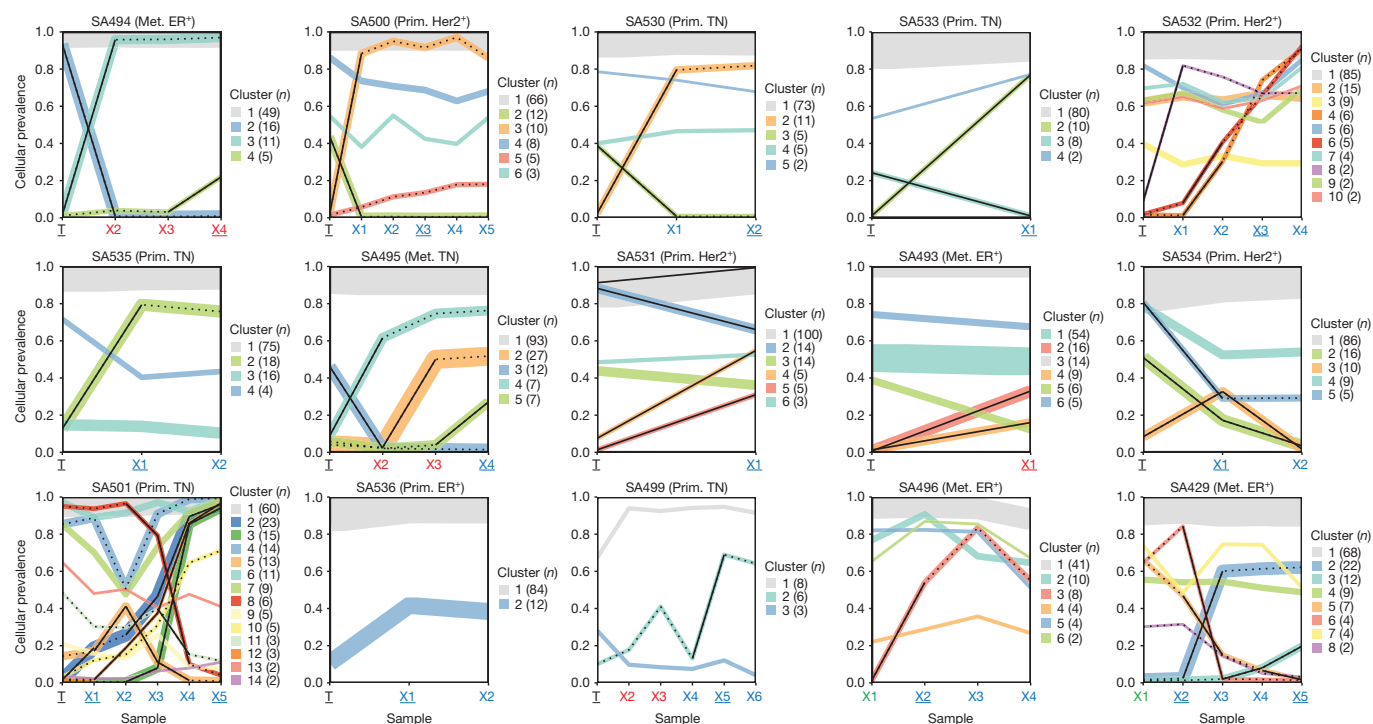


Figure 1 | Clonal dynamics over multiple passages in time. Plots display the mean cellular prevalence estimates of mutation clusters in originating tumours (T) and subsequent xenograft passages (X1, X2, etc.). The clusters and prevalences were inferred by PyClone from bulk population-level targeted deep sequence data. Line widths indicate the number of SNVs comprising each mutation cluster (numbers in brackets adjacent to each plot). Black lines indicate non-neutral dynamics, assessed by non-overlap of credible intervals derived from Bayesian posterior distributions (solid lines, non-neutral over indicated passage; dotted lines, non-neutral over cumulative passages since

initial transplant). All passages that underwent deep sequencing are shown. Transplant sites are represented by colour (blue, subcutaneous; red, subrenal; green, mammary fat pad), tumour and passages analysed by WGSS are underlined. The panels are ordered left to right and top to bottom by the degree of initial change in mutation cellular prevalence. Singleton clusters were not displayed for clarity. ER⁺, oestrogen receptor positive; Her2⁺, Her2 positive; Met., metastatic pleural effusion; Prim., primary breast; TN, triple negative breast cancer.

re-sequencing of SNVs in 210 isolated tumour and xenograft nuclei, using microfluidic devices. We determined evolutionary relationships between nuclei by Bayesian phylogenetic inference²⁰, deriving consensus genotypes for clades representing high probability branch points in the phylogenetic tree.

As predicted by PyClone, two major clades emerge in the SA494 phylogeny, comprising tumour and xenograft nuclei respectively, bearing mutually exclusive sets of alleles in addition to a set of shared alleles (Extended Data Fig. 3a–c and Supplementary Fig. 13). The ancestral clone SNVs (PyClone cluster 1) are common to nuclei from both clades, while SNVs in the predicted dominant tumour clone (cluster 2) and

minor engrafting clone (cluster 3) are restricted to tumour and xenograft nuclei, respectively (Extended Data Fig. 3d, genotypes A and B). This confirms the ancestral relationship between tumour and xenograft, verifies the expansion of a very minor clone (<5%), while also showing unambiguously that mutation clusters inferred by PyClone represent major clonal genotypes.

PyClone analysis of SA501 (Fig. 2 and Supplementary Fig. 12) revealed a dynamic and complex clonal architecture, with gradual expansion of minor mutation clusters observed over consecutive passages, and expansion followed by decline of other clusters (Fig. 2d). The major mutation clusters and their gradual change in prevalence over time

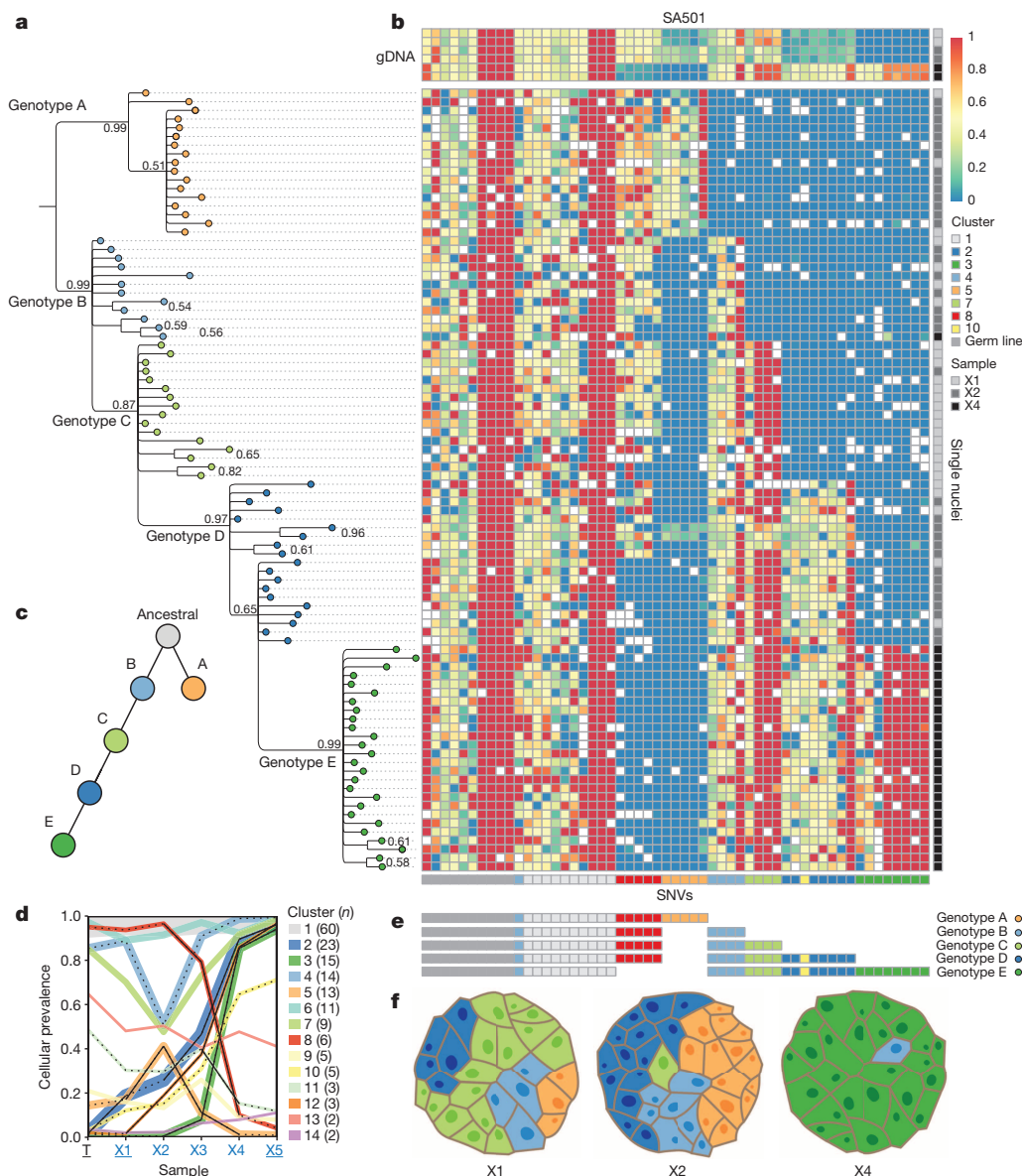


Figure 2 | Single-cell determination of clonal genotypes recapitulates population-based prediction of cascading subclonal evolution. DNA was prepared from 90 individual lysed SA501 xenograft nuclei from passages X1, X2 and X4, and the variant allele ratios were determined by targeted ultra-deep sequencing at 45 somatic SNV and 10 germline SNV positions. **a**, Bayesian phylogenetic tree derived from multi-locus genotypes of individual nuclei, depicting cascading evolution. **b**, Heat map depicting multilocus variant allele ratios (blue/yellow/red corresponds to wild-type/heterozygous/homozygous loci, respectively). Nuclei (y axis) are ordered according to the phylogenetic tree in **a**. gDNA, genomic DNA. Positions (x axis) are grouped according to the consensus genotypes derived from high-probability branch splits in a manner naive to the PyClone clustering. The cluster groupings (horizontal bar below

horizontal axis) recapitulate the PyClone groupings inferred from bulk population measurements (**d**). **c**, Schematic of the phylogeny derived from single-cell genotyping depicts the sequential expansion of genomic subclones. Genotypes are coloured according to the last PyClone mutation cluster acquired at a given point in the phylogeny. **d**, PyClone inference of temporal clonal dynamics from bulk population measurements. **e**, Five consensus genotypes derived from high-probability splits in the phylogenetic tree. **f**, Schematic representations of xenografts X1, X2 and X4 based on single-cell genotypes. Cells are coloured according to their genotype in **c**, and the number of cells within each schematic corresponds to the number of sequenced nuclei with the given genotype in **b**. The relative proportions of cells with each genotype reflect predictions based on bulk measurements in **d**.

predicted by PyClone were confirmed by the clonal genotypes of single cells from SA501 passages X1, X2 and X4 (Fig. 2b and Supplementary Fig. 13). Phylogenetic inference resolved the clonal genotypes of five major clones (Fig. 2a, e), with cascading acquisition of mutations from parental to descendant clone (Fig. 2c). Genotypes A and B belong to sibling clones defined by the addition of cluster 5 and cluster 4 mutations, respectively, to the ancestral genotype defined by clusters 1 and 8; genotype C was derived from genotype B with the addition of mutations in cluster 7; genotype D derived from genotype C with the addition of mutations defined by cluster 2; and genotype E derived from genotype D with the addition of cluster 3 mutations and loss of cluster 8 mutations (Fig. 2a, c, e). The clonal dynamics measured in the population was reflected in the relative abundance of single-cell genotypes in each xenograft (Fig. 2f), mirroring bulk population predictions (Fig. 2d). Both X1- and X2-sampled nuclei show an admixture of clones defined by genotypes A, B, C and D (relatively rare in X1). Genotype E is confined exclusively to X4 nuclei, suggesting that by passage 4, this clone had nearly exhaustively outcompeted its ancestor and sibling clones. Its eventual dominance is mirrored by the decline of genotype A (initially present in X1 and X2), suggesting that the descendants of genotype B outcompeted those of genotype A over time.

Taken together, these single-cell genotyping experiments combined with phylogenetic inference have recapitulated population-level PyClone predictions in a simple (SA494) and a complex (SA501) clonal expansion model. Thus, single-cell genotyping validates PyClone mutation clusters as genomic markers of major clonal genotypes, while providing additional insight into the ancestral lineages of cell populations.

Finally, to determine whether directional clonal dynamics might be associated with deterministic as opposed to stochastic processes (such as random genetic drift), we tested whether similar clonal dynamics occurred when the same tumour population was multiply transplanted into different mice. In 4 of 5 series examined, parallel clonal dynamics of the same mutation cluster(s) were observed (arrows in Fig. 3a, b and Extended Data Fig. 4a, b: SA501 2 of 2 replicate mice at passage X3 and 4 of 4 at X4; SA535 3 of 3 at X1; SA532 3 of 3 at X1, 3 of 7 at X2 and 2 of 2 at X3; SA429 3 of 5 at X2). These include reproducible expansions of initially minor subclones, implying a high likelihood of a shared deterministic mechanism rather than repeated rare stochastic events (for example, arising from transplants close to limiting dilution). In SA501 the same pattern (expansion of cluster 3 mutations mirrored by a decline of cluster 5 mutations) was independently observed in transplants at passage 2, 3 and 4 (2B, 3B and 4A–D in Fig. 3a), suggesting shared clonal fitness but variable timing. We also observed instances of divergence, for example expansion of SA532 cluster 4 specific to branch 1A–2A–3A–4A (Extended Data Fig. 4a). SA535 (Fig. 3b) and SA532 showed examples of clonal expansion patterns replicated in related but different immunodeficient mouse strains (NSG, NRG). To control against shared clonal structure imposed through joint inference of the data sets, we also carried out independent PyClone analyses that excluded all but one transplant at each passage, and observed high correlations of inferred mutation prevalences between same-passage replicates (Extended Data Fig. 5; median Pearson correlations 0.94, 0.93, 0.91, 0.91 and 0.46 for SA501, SA535, SA532, SA429 and SA496, respectively). These data indicate that clonal genotypes defined by somatic aberrations (and/or closely co-segregating genomic factors) can be biologically meaningful determinants of fitness, leading to consistent and reproducible clonal dynamics.

We show here that patient-derived xenograft clonal dynamics on initial transplant vary from polyclonal engraftment with only moderate clonal selection, in which tumour and xenograft clonal prevalence are broadly similar (a minority of cases), to highly skewed dynamics in which initially minor prevalence clones expand to dominate the xenograft (the majority of cases). Expansion of minor subclones has been suggested in previous xenotransplantation studies using malignant epithelial^{10,21–23} or haematopoietic^{24,25} cells, without formal resolution of the clonal genotypes or pattern of subsequent clonal dynamics. In contrast with

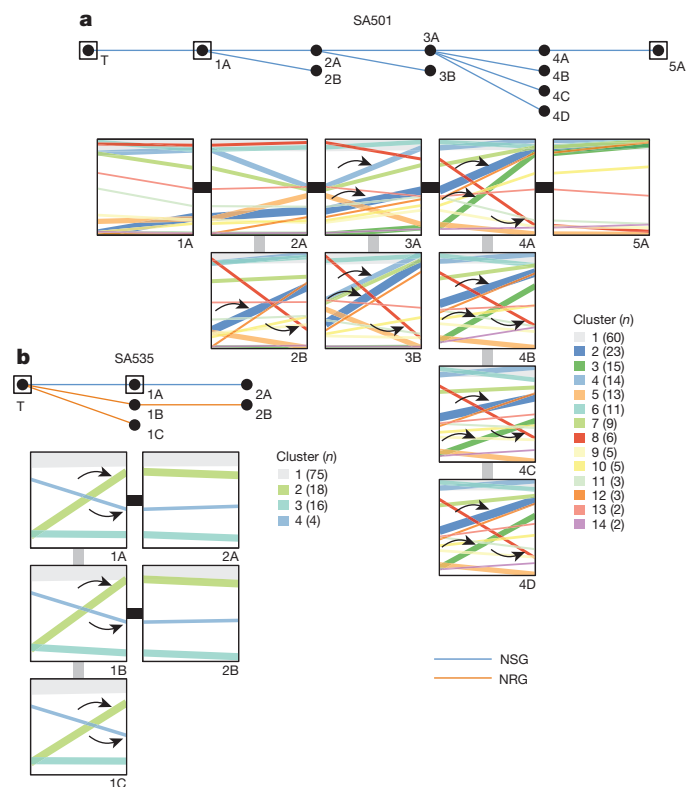


Figure 3 | Clonal dynamics are reproduced in replicate transplants.

a, b, Upper panels, passing history of SA501 and SA535 showing transplants that resulted in successful xenografts. The host mouse strains (blue, NSG; orange, NRG) are indicated. All transplants were in subcutaneous sites. Lower panels, change in cellular prevalence of mutation clusters over individual transplants. Plots correspond to passages in upper panels. The clusters are inferred by PyClone using grouped data from all passages, and correspond to those displayed in Fig. 1. Boxed nodes indicate passages analysed by WGSS. Arrows show examples of parallel clonal dynamics of the same mutation cluster in multiple replicate transplants.

preliminary studies of xenoengraftment, we find correlated dynamics of clones defined by SNVs or copy-number aberrations as clonal marks. Expansion patterns are most often pronounced in the initial establishment passage; however, in cases where initial clonal selection is weak, subsequent evolution over passaging is more evident. Furthermore, polyclonal sub-structure may emerge even in xenografts that have undergone a modest population bottleneck on initial engraftment. These dynamic processes are not evident from histopathological or imaging characteristics, which remain broadly stable, consistent with previous reports^{8,9,23}.

Notably, we find that the population dynamics of genomically defined clones are replicated when transplants are carried out in multiple mice, implying that the basis of selection is non-random and probably closely linked to the particular mutation genotype (or epigenotype) that defines the clone. The most parsimonious explanation for repeated observation of these clonal dynamics is that the clones are mostly pre-existing, and variations in clonal fitness explain the dynamic behaviour, as opposed to *de novo* somatic mutation. Furthermore, cases in which conversion from minor to dominant clone occurs monotonically over multiple passages demonstrate that selective fitness can be persistent rather than transient. Thus, specific somatic genotypes are likely to act as genetic markers of clonal growth and fitness advantages, yielding predictable and reproducible clonal dynamics. Determination of the precise aberrations that give rise to selective clonal fitness still faces considerable challenges. In this regard, we believe that ascertainment of clonal dynamics will prove essential for fully informed future studies of drug response and tumour biology in xenografts of human breast cancers.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 22 August 2013; accepted 8 October 2014.

Published online 26 November 2014.

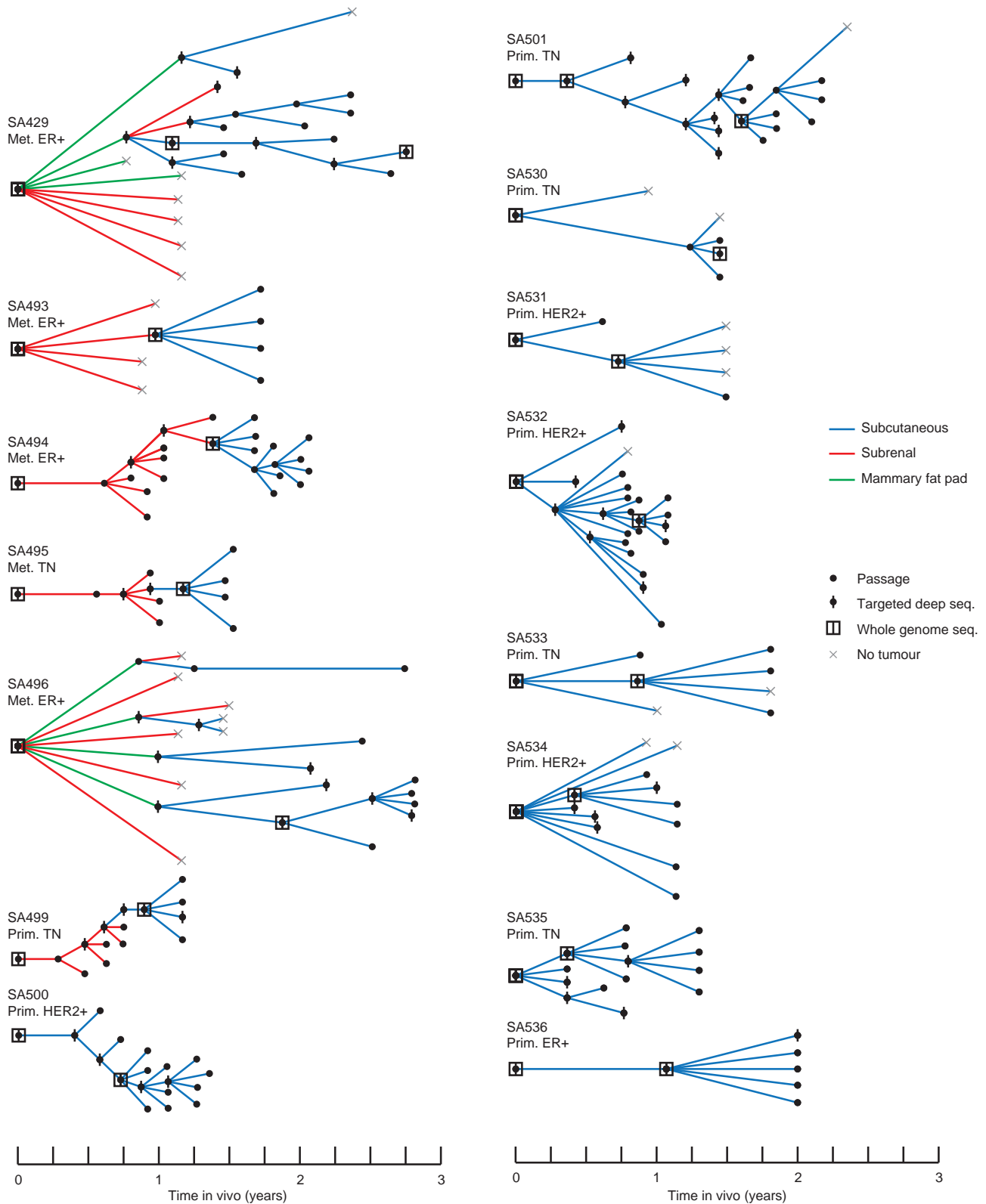
1. Aparicio, S. & Caldas, C. The implications of clonal genome evolution for cancer medicine. *N. Engl. J. Med.* **368**, 842–851 (2013).
2. Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194**, 23–28 (1976).
3. Diaz, L. A. Jr *et al.* The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers. *Nature* **486**, 537–540 (2012).
4. Shah, S. P. *et al.* The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* **486**, 395–399 (2012).
5. Gerlinger, M. *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* **366**, 883–892 (2012).
6. Campbell, P. J. *et al.* The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature* **467**, 1109–1113 (2010).
7. Bashashati, A. *et al.* Distinct evolutionary trajectories of primary high-grade serous ovarian cancers revealed through spatial mutational profiling. *J. Pathol.* **231**, 21–34 (2013).
8. DeRose, Y. S. *et al.* Tumor grafts derived from women with breast cancer authentically reflect tumor pathology, growth, metastasis and disease outcomes. *Nature Med.* **17**, 1514–1520 (2011).
9. Zhang, X. *et al.* A renewable tissue resource of phenotypically stable, biologically and ethnically diverse, patient-derived human breast cancer xenograft models. *Cancer Res.* **73**, 4885–4897 (2013).
10. Ding, L. *et al.* Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* **464**, 999–1005 (2010).
11. Pearson, T. *et al.* Non-obese diabetic-recombination activating gene-1 (NOD-*Rag1^{null}*) interleukin (IL)-2 receptor common gamma chain (*IL2 γ ^{null}*) null mice: a radioresistant model for human lymphohaematopoietic engraftment. *Clin. Exp. Immunol.* **154**, 270–284 (2008).
12. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
13. Ha, G. *et al.* Integrative analysis of genome-wide loss of heterozygosity and monoallelic expression at nucleotide resolution reveals disrupted pathways in triple-negative breast cancer. *Genome Res.* **22**, 1995–2007 (2012).
14. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
15. Ellis, M. J. *et al.* Whole-genome analysis informs breast cancer response to aromatase inhibition. *Nature* **486**, 353–360 (2012).
16. Banerji, S. *et al.* Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature* **486**, 405–409 (2012).
17. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).
18. Roth, A. *et al.* PyClone: statistical inference of clonal population structure in cancer. *Nature Methods* **11**, 396–398 (2014).
19. Ha, G. *et al.* Titan: inference of copy number architectures in clonal cell populations from tumor whole genome sequence data. **24**, 1881–1893 (2014).
20. Ronquist, F. *et al.* MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61**, 539–542 (2012).
21. Kreso, A. *et al.* Variable clonal repopulation dynamics influence chemotherapy response in colorectal cancer. *Science* **339**, 543–548 (2013).
22. Nolan-Steva, O. *et al.* Measurement of cancer cell growth heterogeneity through lentiviral barcoding identifies clonal dominance as a characteristic of *in vivo* tumor engraftment. *PLoS ONE* **8**, e67316 (2013).
23. Li, S. *et al.* Endocrine-therapy-resistant *ESR1* variants revealed by genomic characterization of breast-cancer-derived xenografts. *Cell Rep.* **4**, 1116–1130 (2013).
24. Notta, F. *et al.* Evolution of human *BCR-ABL1* lymphoblastic leukaemia-initiating cells. *Nature* **469**, 362–367 (2011).
25. Clappier, E. *et al.* Clonal selection in xenografted human T cell acute lymphoblastic leukemia recapitulates gain of malignancy at relapse. *J. Exp. Med.* **208**, 653–661 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements We are grateful to the staff of the CTAG Molecular Pathology facility, members of the Library Technical Development, Library Construction, Sequencing and Bioinformatics teams at the Michael Smith Genome Sciences Centre for technical assistance with data generation, and S. Kaloger for assistance with sample collection. S.A. and S.P.S. are supported by Canada Research Chairs. P.E. is supported by a Michael Smith Foundation for Health Research (MSFHR) Fellowship. A.S. is supported by an NSERC CREATE scholarship through the graduate program in Genome Science and Technology at UBC. S.P.S. is a MSFHR scholar. We acknowledge long-term funding support provided by the BC Cancer Foundation. The S.A., S.P.S. and C.H. groups receive operating funds from the Canadian Breast Cancer Foundation, Canadian Cancer Society Research Institute, Terry Fox Research Institute, Genome Canada and Canadian Institutes for Health Research (CIHR). We thank S. Mullaly for critical reading of the manuscript.

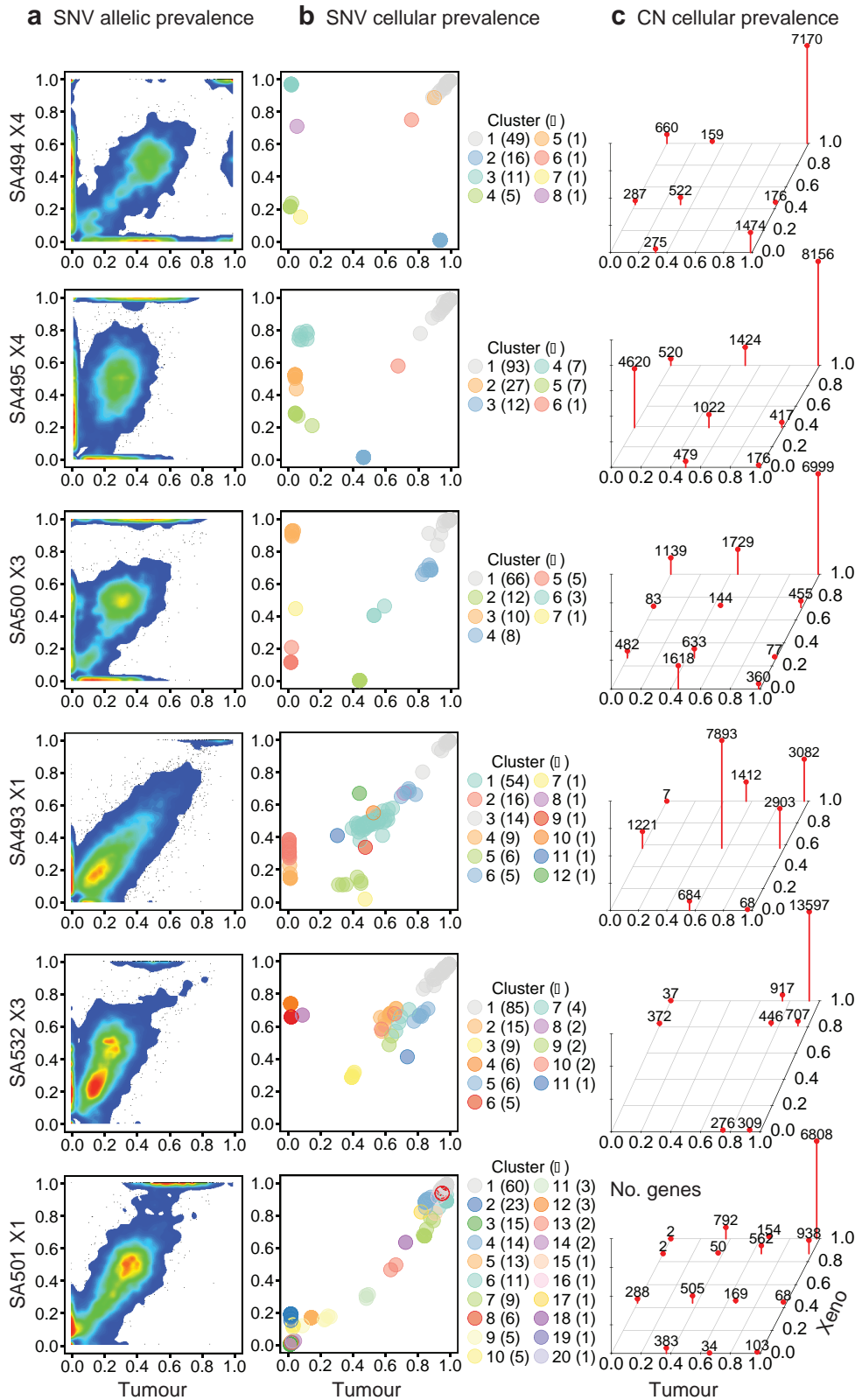
Author Contributions S.A. and S.P.S. designed the study and supervised the research. S.A., S.P.S., P.E. and A.S. wrote the paper. P.E., A.Br., J.S., T.A., W.G., H.C., H.X., L.N., Y.W. and D.L. performed transplants and passaging. K.G., S.C. and C.M. recruited patients and performed tissue biopsies. A.S., P.E., G.H., C.N., H.F., A.J.L.R., C.L., A.Ba., C.S., K.S., J.R., R.A., A.M., C.d.S., S.P.S. and S.A. carried out bioinformatics analyses. J.K., D.Y., E.L., J.Br., A.W., J.Bi., K.L., A.J.M., A.O., R.M., Y.Z., C.H. and M.A.M. assisted with sequence generations and single-cell experiments. T.O., J.L. and D.H. contributed to histological analysis. C.J.E., C.H., M.A.M., C.C., S.P.S. and S.A. provided intellectual contributions to design or interpretation.

Author Information Genome data have been deposited at the European Genome-phenome Archive (<http://www.ebi.ac.uk/ega>) under accession number EGAS00001000952. Processed data can be viewed at <http://www.cbioportal.org>. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.A. (sapario@bccrc.ca) or S.P.S. (sshah@bccrc.ca).



Extended Data Figure 1 | Transplant history. Diagrams show the transplant history of each xenograft line. Line segment colours represent the site used for each transplant (blue, subcutaneous; red, subrenal capsule; green, mammary fat pad). Black points indicate the passage of an engrafted xenograft to the next mouse generation. Grey crosses indicate transplants that did not result in palpable tumours. Samples analysed by whole-genome and/or targeted deep

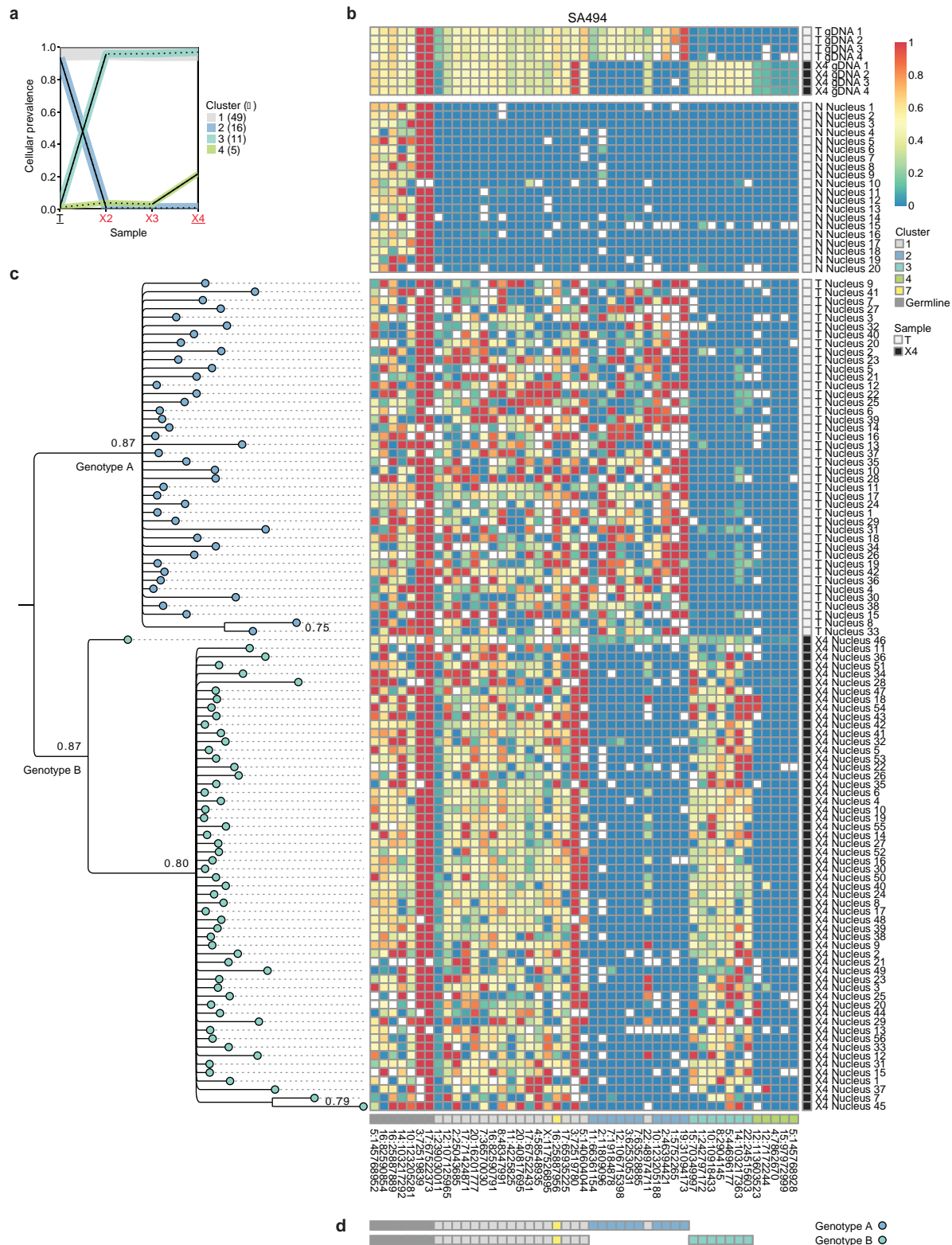
sequencing are indicated (black squares and vertical lines, respectively). The cumulative time *in vivo* is shown on the x axis. The originating tumour site (Met., pleural effusion; Prim., primary breast) and immunohistochemical expression of biomarkers (ER, oestrogen receptor; PR, progesterone receptor; TN, triple negative for ER, PR and HER2) are shown.



Extended Data Figure 2 | Comparison of the prevalence of mutations in six originating tumours and subsequent xenografts in SNV and CNA spaces.

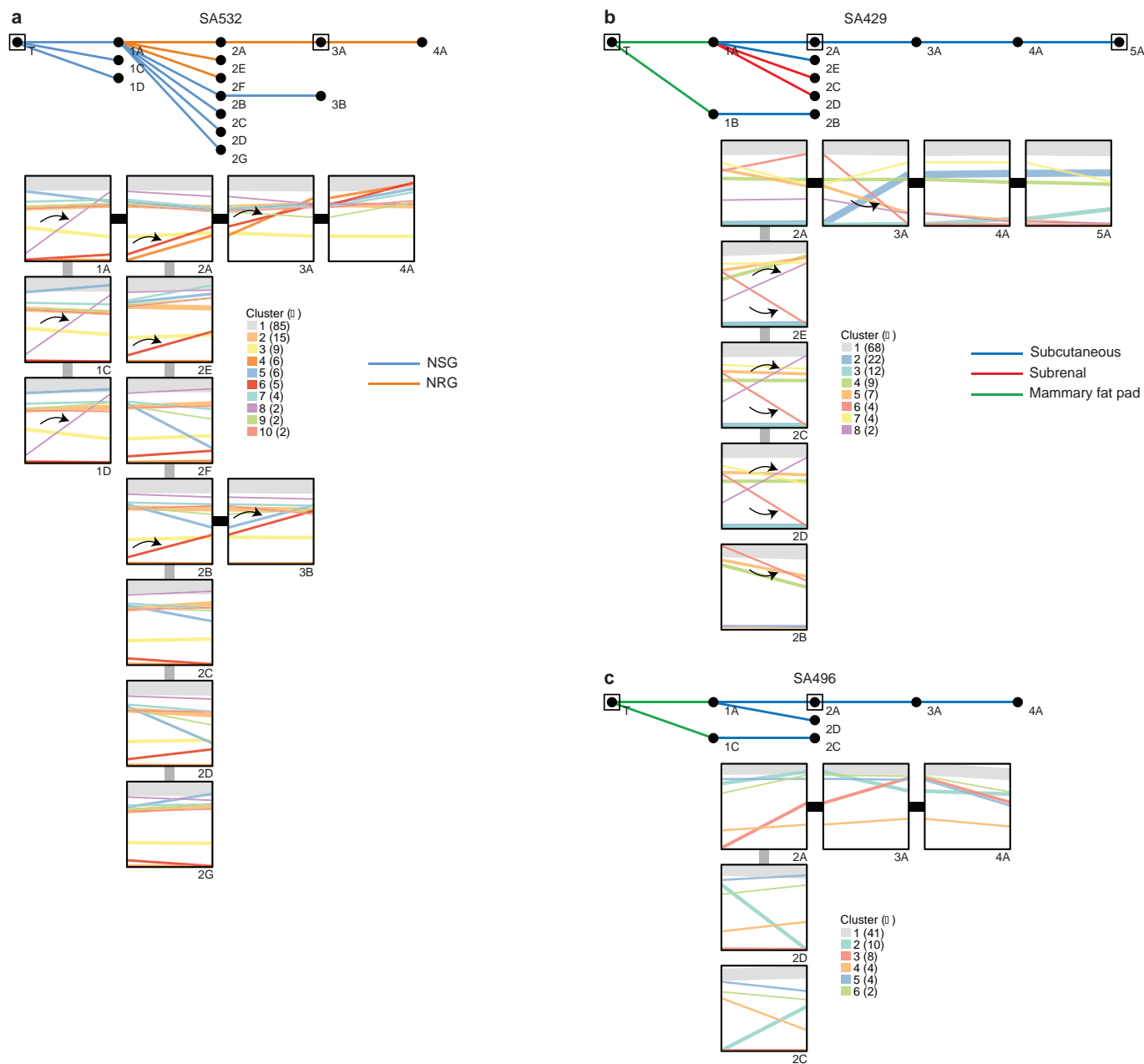
a, Density scatter plots showing the WGSS variant allelic prevalence of genome-wide high-confidence SNVs in tumours (x axis) and xenografts (y axis). SNVs in clones undergoing neutral dynamics lie along a diagonal, and SNVs in clones undergoing expansion or contraction lie on/towards the y and x axes, respectively. **b**, Scatter plots showing the mutation cellular prevalence of selected SNVs in tumours and xenografts, inferred by PyClone from

population-level targeted deep sequencing. Circles represent individual SNVs, colours indicate clusters of mutations for which mutation cellular prevalences vary together over all sample time points. **c**, Scatter plots show co-occurrence of CNA/LOH events inferred by TITAN in tumours and xenografts. The z axis height of each bar shows the number of genes belonging to a unique mutation cluster and present at the indicated mutation cellular prevalence in tumour (x axis) and xenograft (y axis).



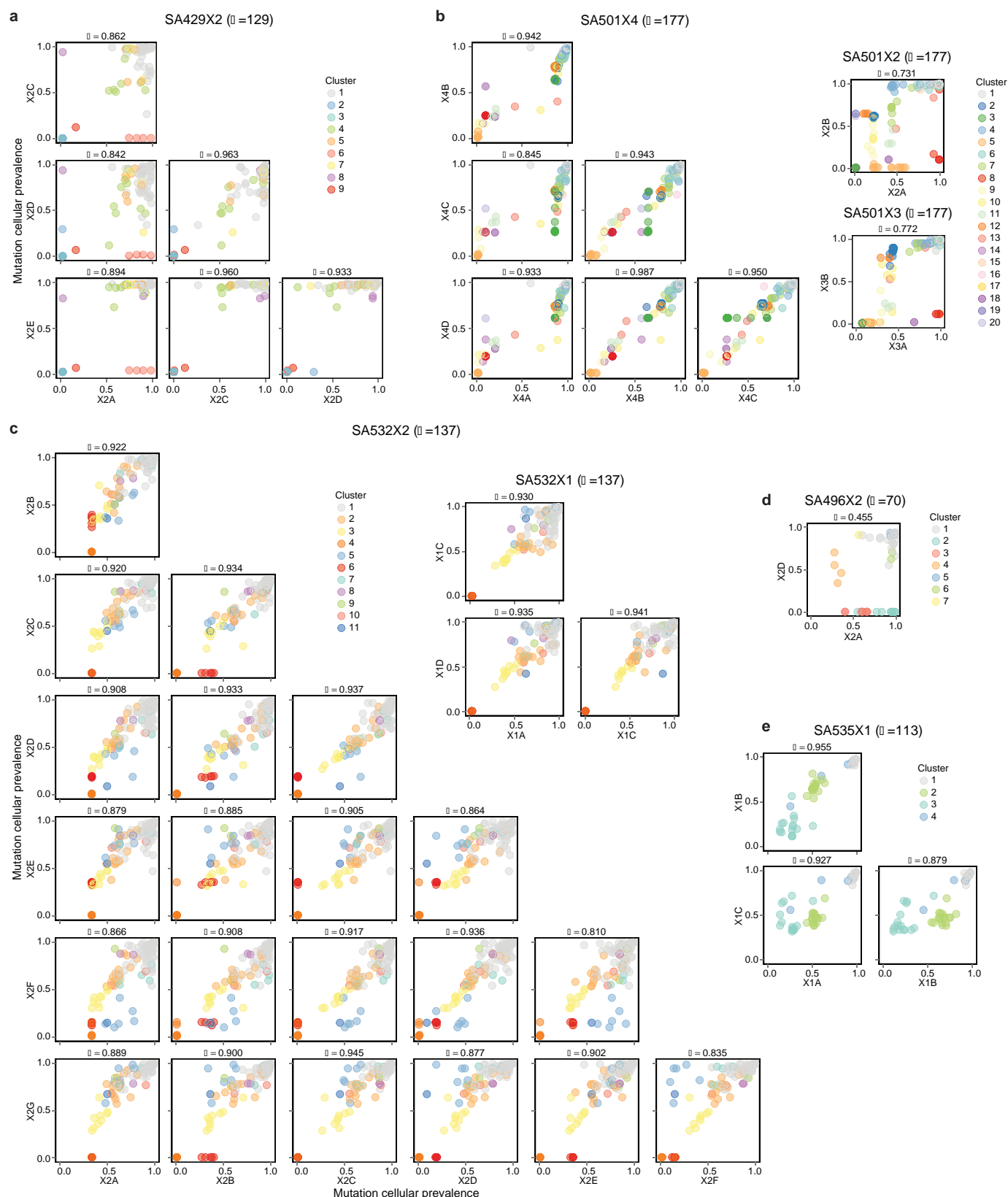
Extended Data Figure 3 | Single-cell determination of clonal genotypes recapitulates population-based prediction of minor clone selection. DNA prepared from 62 individual lysed SA494 tumour and 58 passage 4 lysed xenograft nuclei was amplified in single reactions using a panel of multiplexed PCR primer pairs targeting amplicons containing 40 SNV and 7 germline variants, and the variant allele ratios were determined by targeted deep sequencing. **a**, Mutation clusters inferred by the PyClone model from bulk population measurements. **b**, Bayesian phylogenetic tree derived from multi-locus genotypes of individual nuclei. The tumour and xenograft nuclei group in distinct clades. **c**, Heat map depicts the multi-locus variant allele prevalences

(blue/yellow/red corresponds to wild-type/heterozygous/homozygous loci, respectively) at variant positions (horizontal axis) in individual nuclei (vertical axis, ordered by phylogenetic grouping in **b**). Upper two blocks show gDNA controls and normal cell nuclei present in tumour samples. The PyClone mutation cluster corresponding to each SNV is indicated by colour in the lowermost horizontal bar. **d**, Consensus genotypes derived from high-probability splits in the phylogenetic tree confirm a set of high prevalence tumour-specific and xenograft-specific mutations, consistent with the expansion of a minor originating clone to dominance in the xenograft, as well as mutations shared in tumour and xenograft nuclei.



Extended Data Figure 4 | Clonal dynamics are reproduced in replicate transplants. **a–c**, Upper panels, passing history of SA532, SA429 and SA496, showing transplants that resulted in successful xenografts. The transplant sites (blue, subcutaneous; red, subrenal; green, mammary fat pad; all subcutaneous for SA532) and host mouse strains (blue, NSG; orange, NRG; all NSG for SA429 and SA496) are shown. Boxed nodes indicate passages analysed by WGSS. Lower panels, change in cellular prevalence of mutation

clusters over individual transplants. Plots correspond to passages in upper panels. The clusters are inferred by PyClone using grouped data from all passages and correspond to those displayed in Fig. 1. Arrows in SA429 and SA532 show examples of parallel clonal dynamics of the same mutation cluster in multiple replicate transplants. SA496 exhibits less replicated evolution compared with other cases.



Extended Data Figure 5 | Correlation of clonal dynamics in replicate transplants of SA429, SA501, SA532, SA496 and SA535. a–e, Scatter plots display the inferred mutation cellular prevalence of all SNVs in pairs of same-passage replicates, for cases SA429, SA501, SA532, SA496 and SA535, respectively. For each replicate, prevalences are inferred by a separate PyClone

analysis that excludes data from other same-passage transplants. Colours indicate mutation clusters inferred in each individual PyClone analyses; the SNVs clustered and colours assigned may differ in each plot. The Pearson correlation coefficients are shown, indicating closely related evolution in most pairs.

Catalysts from synthetic genetic polymers

Alexander I. Taylor¹, Vitor B. Pinheiro^{1†}, Matthew J. Smola², Alexey S. Morgunov¹, Sew Peak-Chew¹, Christopher Cozens¹, Kevin M. Weeks², Piet Herdewijn^{3,4} & Philipp Holliger¹

The emergence of catalysis in early genetic polymers such as RNA is considered a key transition in the origin of life¹, pre-dating the appearance of protein enzymes. DNA also demonstrates the capacity to fold into three-dimensional structures and form catalysts *in vitro*². However, to what degree these natural biopolymers comprise functionally privileged chemical scaffolds³ for folding or the evolution of catalysis is not known. The ability of synthetic genetic polymers (XNAs) with alternative backbone chemistries not found in nature to fold into defined structures and bind ligands⁴ raises the possibility that these too might be capable of forming catalysts (XNAzymes). Here we report the discovery of such XNAzymes, elaborated in four different chemistries (arabino nucleic acids, ANA⁵; 2'-fluoroarabino nucleic acids, FANA⁶; hexitol nucleic acids, HNA; and cyclohexene nucleic acids, CeNA⁷) directly from random XNA oligomer pools, exhibiting in *trans* RNA endonuclease and ligase activities. We also describe an XNA-XNA ligase metalloenzyme in the FANA framework, establishing catalysis in an entirely synthetic system and enabling the synthesis of FANA oligomers and an active RNA endonuclease FANAzyme from its constituent parts. These results extend catalysis beyond biopolymers and establish technologies for the discovery of catalysts in a wide range of polymer scaffolds not found in nature⁸. Evolution of catalysis independent of any natural polymer has implications for the definition of chemical boundary conditions for the emergence of life on Earth and elsewhere in the Universe⁹.

Life is dependent on catalysis, as many chemical transformations essential for cellular function are kinetically sluggish and/or thermodynamically disfavoured under ambient conditions. The emergence of a catalyst (or catalytic system) for RNA self-replication is considered to have been a key event in the origin of life. Thus the development of molecular heredity itself depends not only on the capacity of nucleic acids for genetic information storage and retrieval but also on their ability to form catalysts¹. Although proteins have largely supplanted this role in present-day biology, nucleic-acid-mediated catalysis remains crucial, notably in RNA processing¹⁰ and translation¹¹. Furthermore, a range of RNA and DNA enzymes (ribozymes/DNAzymes) have been discovered by *in vitro* evolution¹².

Catalysis by nucleic acids (and by biopolymers in general) requires as a minimum the presence of chemically functional groups and a framework for their precise arrangement. Synthetic genetic polymers (XNAs) with backbones based on congeners of the canonical ribofuranose share with RNA and DNA a capacity for heredity, evolution and the ability to fold into defined three-dimensional structures, forming ligands (aptamers)⁴. We therefore sought to establish whether XNAs could also support the evolution of catalysts.

Taking advantage of XNA replication technology developed previously⁴, we devised a strategy for the discovery of RNA endonuclease XNAzymes by cleavage of an internal RNA sequence (Extended Data Fig. 1). Chimeric RNA-XNA libraries were prepared by RNA-primed XNA synthesis in four scaffolds using mutant polymerases: D4K⁴ for arabinonucleic acid (ANA)⁵ and fluoro-arabinonucleic acid (FANA)¹³, 6G12⁴ for cyclohexenyl nucleic acid (CeNA)⁷ and a newly engineered 6G12 I521L variant (see Methods) for 1,5 anhydrohexitol nucleic acid (HNA)⁷. After 13–17

rounds of selection, polyclonal pools showed RNA endonuclease activity and were deep sequenced (Extended Data Fig. 2). Abundant sequences across all four XNAs were tested for intramolecular (in *cis*) endonuclease activity and a subset of active clones for bimolecular (in *trans*) activity. We further examined one RNA endonuclease XNAzyme for each scaffold (FR17_6 (FANA), AR17_5 (ANA), HR16_1 (HNA) and CeR16_3 (CeNA)) (Fig. 1). All showed site-specific sequence-dependent (Extended Data Fig. 3) RNA cleavage with a range of catalytic rates ($k_{\text{obs}} = 0.06\text{--}0.0001\text{ min}^{-1}$ at 25 °C). While the rate of the FR17_6 XNAzyme is comparable to analogous ribozymes and DNAzymes, ANA and in particular HNA and CeNA catalysts are 20–600-fold slower. Nevertheless, all four catalyse RNA cleavage through a classic transesterification mechanism (as seen in, for example, the 'hammerhead' or 'hairpin' ribozymes¹⁴), yielding products with 2',3' cyclic phosphate and 5' hydroxyl groups (Extended Data Fig. 4).

We dissected contributions of individual nucleotides in the FR17_6 XNAzyme, defining a 26 nucleotide (nt) catalytic core (FR17_6min). As all four FANA nucleotide phosphoramidites are commercially available, this minimized XNAzyme could be prepared by solid-phase synthesis (see Methods) and was found to retain near full activity (Fig. 2a–c; $k_{\text{obs}} = 0.026\text{ min}^{-1}$ at 25 °C), including multiple turnover catalysis (Fig. 2d). FR17_6min shows a pH optimum (pH_{opt}) of 9.25 (Extended Data Fig. 4h), consistent with a mechanism involving deprotonation of the cleavage site-proximal 2' hydroxyl. A screen of Irving-Williams divalent metals reveals that FR17_6min is Mg^{2+} -dependent with an apparent Michaelis constant $K_{\text{m}} \approx 30\text{ mM}$ (Extended Data Fig. 4i), with only Mn^{2+} able to partially restore activity (Extended Data Fig. 4g).

The secondary structure of FR17_6, including an inert RNA substrate modified with 2'-O-Me at the cleavage site (Extended Data Fig. 5), was probed by selective 2' hydroxyl acylation analysed by primer extension (SHAPE)¹⁵ (for RNA, modifying 2' OH at flexible regions) and/or dimethyl sulphate (DMS)¹⁶ (for FANA, modifying primarily unpaired adenine and cytosine). This secondary structure is broadly similar to other RNA-acting nucleic acid catalysts, with a central domain flanked by substrate-binding arms (P1, P2), albeit with a 3 nt bulge in P2 (Extended Data Fig. 5c).

In general, the RNA endonuclease XNAzymes are novel sequences, although some in the ANA system (Extended Data Fig. 2c) retain partial sequences from the 8-17 and 10-23 DNAzymes¹⁷ used in library design (in addition to N₄₀ sequences) (see Methods). The AR17_5 ANAzyme shares 12 of the 14 core residues of the 8-17 DNAzyme, as well as A'G > N'G cleavage preference (Extended Data Fig. 3b). However, conversion of the complete 8-17 sequence into ANA (or indeed FANA, HNA or CeNA) yields no activity (Extended Data Fig. 3e), indicating the acquisition or rearrangement of key residues during selection. Nevertheless, topological similarities (without sequence homology) between FR17_6 and this family of DNAzymes¹⁸ suggests the possibility that for XNAs that form DNA-like B-form duplexes, such as ANA and FANA (albeit with a non-canonical O4'-endo (east) sugar conformation)¹⁹, catalysts may reside in the structural or sequence vicinity of extant DNAzymes.

¹MRC Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge Biomedical Campus, Cambridge CB2 0QH, UK. ²Department of Chemistry, University of North Carolina, Chapel Hill, North Carolina 27599-3290, USA. ³KU Leuven, Rega Institute, Minderbroedersstraat 10, B 3000 Leuven, Belgium. ⁴Université Evry, Institute of Systems and Synthetic Biology, 5 rue Henri Desbrières, 91030 Evry Cedex, France. [†]Present address: Institute of Structural and Molecular Biology, Department of Biological Sciences, Birkbeck, University of London, Malet Street, London WC1E 7HX, UK, and Structural and Molecular Biology Department, University College London, Gower Street, London WC1E 6BT, UK.

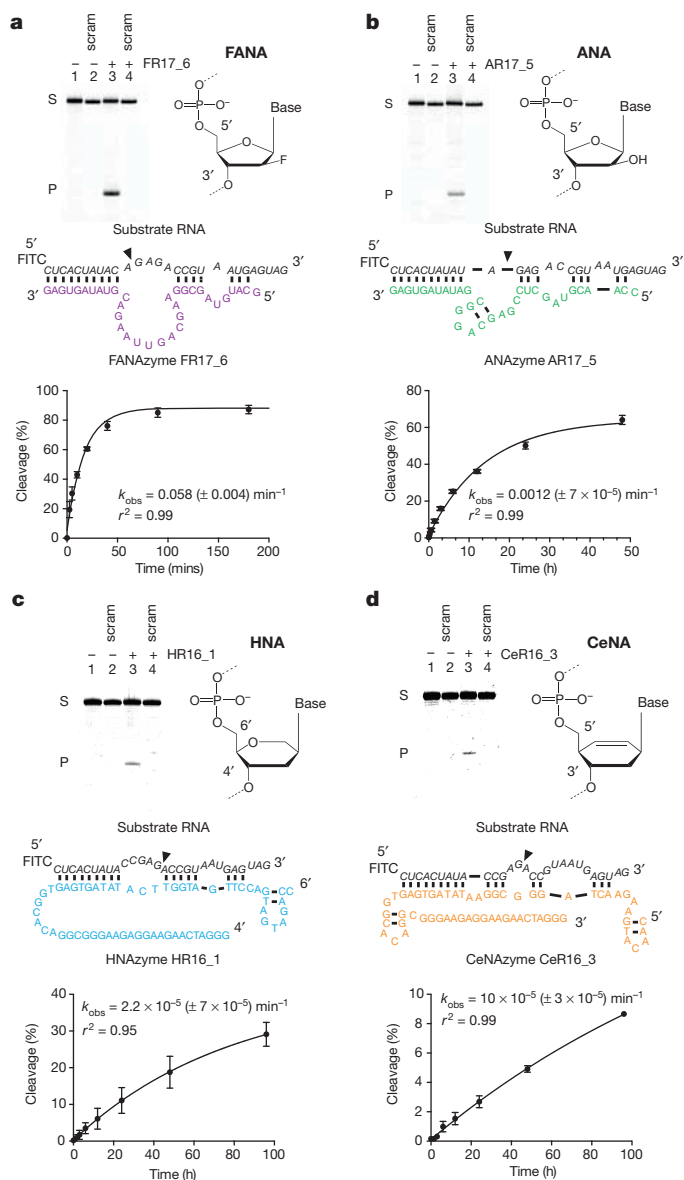


Figure 1 | RNA endonuclease XNAzymes elaborated in four synthetic genetic polymer chemistries. Shown are gel electrophoretograms, putative secondary structures, and pre-steady state reaction rates (k_{obs}) at 25 °C ($n = 3$; error bars, s.d.) of enzymes composed of XNA: FANA (a), ANA (b), HNA (c) and CeNA (d). Urea-PAGE gels show bimolecular cleavage (in *trans*) of cognate RNA substrates ($\text{NucS}^{\text{R}}_{\text{min}}$ variants, see Extended Data Fig. 3) (lanes 1 and 3), but not scrambled RNA ($\text{NucS}^{\text{R}}_{\text{scram}}$) (lanes 2 and 4), catalysed by XNAzymes (lanes 3 and 4). Bands representing substrates and products are marked S and P, respectively.

Having established the capacity for catalysis in four different XNA backbones, we wondered whether XNAzymes could be evolved to ligate RNA as well as cleave it^{20–22}. We selected for RNA–RNA ligase activity using a bi-molecular strategy: 5'-RNA–XNA libraries carrying 5' triphosphate moieties (5' ppp)(Extended Data Fig. 6) were challenged to ligate to DNA–RNA–3' substrates. We identified RNA ligase XNAzymes (FANA) by deep sequencing and screening (Extended Data Fig. 7), and chose one (F2R17_1) for further characterization. A minimized (39 nt), chemically synthesized version (F2R17_1min) was capable of ligating two RNA substrates (LigS1^{R} (3'-OH) + LigS2^{R} (5' ppp)) in a tri-molecular reaction (Fig. 3) with 'natural' regioselectivity (3'-5' rather than 2'-5'), as judged by comparison with 'mock' RNA using strong anion exchange chromatography (SAX-HPLC)(Extended Data Fig. 7c). The reaction rate is low ($k_{\text{obs}} = 2 \times 10^{-4} \text{ min}^{-1}$ at 25 °C), but represents an enhancement ($k_{\text{obs}}/k_{\text{uncat}}$) of

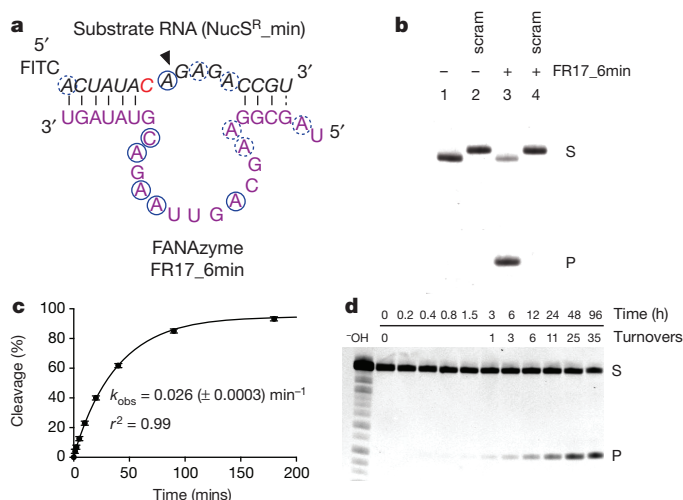


Figure 2 | Chemical synthesis yields an active RNA endonuclease XNAzyme. a, Secondary structure of truncated FANAzyme FR17_6 (FR17_6min, purple), determined by RNA-SHAPE and RNA/FANA-DMS mapping (derived from a larger construct, Extended Data Fig. 5). Red indicates SHAPE reactive residues (RNA), blue circles indicate DMS-reactive (preferentially A or C) residues (RNA or FANA), dashed circles indicate marginal reactivity. The substrate RNA ($\text{NucS}^{\text{R}}_{\text{min}}$) is a minimized version of the substrate for FR17_6. b, FR17_6min synthesized using FANA phosphoramidites (Extended Data Fig. 10) cleaves cognate RNA substrate ($\text{NucS}^{\text{R}}_{\text{min}}$; lanes 1 and 3), but not a scrambled RNA ($\text{NucS}^{\text{R}}_{\text{scram}}$; lanes 2 and 4), with c, essentially unchanged catalytic rate (k_{obs}) at 25 °C ($n = 3$; error bars, s.d.). d, FR17_6min (10 nM) can perform multiple turnover cleavage of RNA $\text{NucS}^{\text{R}}_{\text{min}}$ (1 μM).

10^4 -fold compared to the uncatalysed background reaction (RNA substrates LigS1^{R} + LigS2^{R} hybridized to a 'splint' (complementary FANA template); $k_{\text{uncat}} = 2 \times 10^{-8} \text{ min}^{-1}$). Like the RNA endonuclease FANAzyme FR17_6, the activity of F2R17_1min is enhanced at basic pH ($\text{pH}_{\text{opt}} = 10.25$; Extended Data Fig. 7e), as well as by Mg^{2+} (Extended Data Fig. 7f), for which only Mn^{2+} can be substituted (Extended Data Fig. 7d), consistent with a mechanism involving deprotonation and nucleophilic attack of the 3' hydroxyl of LigS1^{R} on the α -phosphate of 5' ppp- LigS2^{R} , analogous to, for example, the R3C ligase ribozyme²³.

In the above examples, XNA catalysts cleave or ligate natural substrates (RNA, DNA). Next we sought to discover whether XNA catalysts could act on XNA substrates, establishing a fully synthetic catalytic system. We chose to select for XNA–XNA ligase activity with a view to its potential synthetic utility for the assembly of larger XNA oligomers (Extended Data Fig. 8). Again exploiting solid-phase FANA synthesis for substrate and primer strands, we synthesized an all-FANA library loosely patterned on the secondary structure of the DNA-ligase DNAzyme E47²⁴ (see Methods) and selected for ligation of the library 5' hydroxyl group to a substrate activated with 3' phosphorylimidazolide (pIm). After 4 rounds, we identified multiple FANA ligase FANAzymes (Extended Data Fig. 9). One of these, FpImR4_2 (41nt), was found to be a Zn^{2+} -dependent metalloenzyme capable of XNA–XNA (FANA–FANA) ligation in a trimolecular reaction ($\text{LigS1}^{\text{F}} + \text{LigS2}^{\text{F}} + \text{FpImR4}_2 \rightarrow \text{LigP}^{\text{F}} + \text{FpImR4}_2$; Fig. 4). The product (LigP^{F}) shows an identical SAX-HPLC profile to a 'mock' product synthesized by D4K polymerase (Extended Data Fig. 9c), suggesting the ligation proceeds with 3'-5' regioselectivity. Despite the higher reactivity of the activating group, uncatalysed FANA reactions with or without a complementary FANA splint yielded no detectable ligation of LigS1^{F} and LigS2^{F} (Fig. 4b), even after incubation for several days.

Despite no apparent sequence or structural homology, as judged by DMS probing (Extended Data Fig. 5), FpImR4_2 and DNAzyme E47 may employ analogous catalytic strategies because they display a similar pH optimum ($\text{pH}_{\text{opt}} = 7.25$; Extended Data Fig. 9e), metal ion dependence (Zn^{2+} ; Extended Data Fig. 9f) and catalytic rate (FpImR4_2

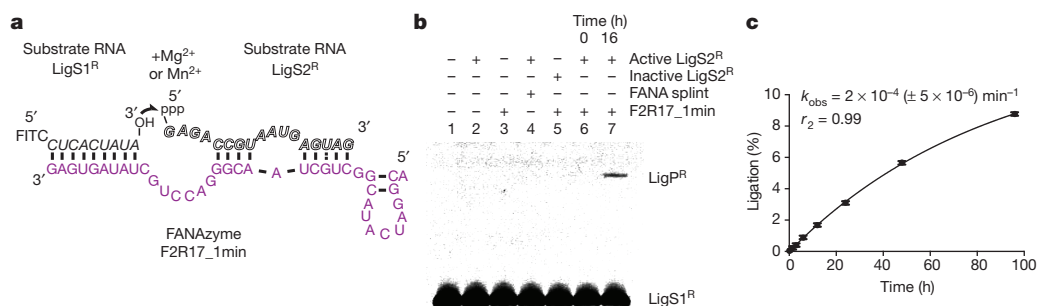


Figure 3 | An RNA ligase XNAzyme (FANA). **a**, Putative secondary structure of truncated chemically synthesized FANAzyme (F2R17_1min, purple) that ligates RNA substrate LigS1^R to LigS2^R, activated with 5' triphosphate (ppp), in a trimolecular reaction in *trans*. **b**, Urea-PAGE gel showing no significant product (LigP^R) observed with: substrate LigS1^R alone (lane 1), no XNAzyme (lane 2), no LigS2^R (lane 3), complementary FANA splint (lane 4), or LigS2^R

lacking 5' ppp (lane 5); product formation is dependent on LigS1^R, activated LigS2^R and XNAzyme (lanes 6 and 7). No product was detectable with combinations of RNA, DNA or FANA versions of LigS1 and (5'ppp)LigS2, except DNA LigS1 and RNA LigS2, which showed ~1.5% ligation after 20 h (Extended Data Fig. 7g). **c**, Pre-steady state trimolecular reaction rate (k_{obs}) at 25 °C ($n = 3$; error bars, s.d.).

$k_{\text{obs}} = 0.04 \text{ min}^{-1}$ versus E47 $k_{\text{obs}} = 0.06 \text{ min}^{-1}$ at 35 °C²⁴. However, unlike with E47, in the FpImR4_2 reaction, Cu²⁺ cannot substitute Zn²⁺, and Mg²⁺ (or Ca²⁺) enhances activity (Extended Data Fig. 9d). Unlike RNA ligase FANAzyme F2R17_1, FpImR4_2 displays a relaxed recognition of substrate chemistry; although most efficient at FANA-FANA ligation, it can also ligate FANA-DNA, FANA-RNA, DNA-FANA as well as DNA-DNA (Extended Data Fig. 9g).

Finally, in order to explore the synthetic potential of XNA ligation, the substrate strands and the XNA ligase FANAzyme were adapted to

perform novel reactions. Modification of LigS2^F substrate to include the LigS1^F sequence (that is, LigS2+1^F) and a 3'-phosphorylimidazole activation group enabled iterative substrate addition, thus synthesizing FANA oligomers up to 100 nt long (Fig. 4d). Modification of the FpImR4_2 substrate-binding strands allowed ligation of a variant of the FR17-6 RNA endonuclease from constituent fragments (see Methods), enabling XNAzyme-catalysed synthesis of another XNAzyme (Fig. 4e).

Synthesis, replication (via a DNA intermediate) and evolution of synthetic genetic polymers (XNAs) not found in nature has opened up new

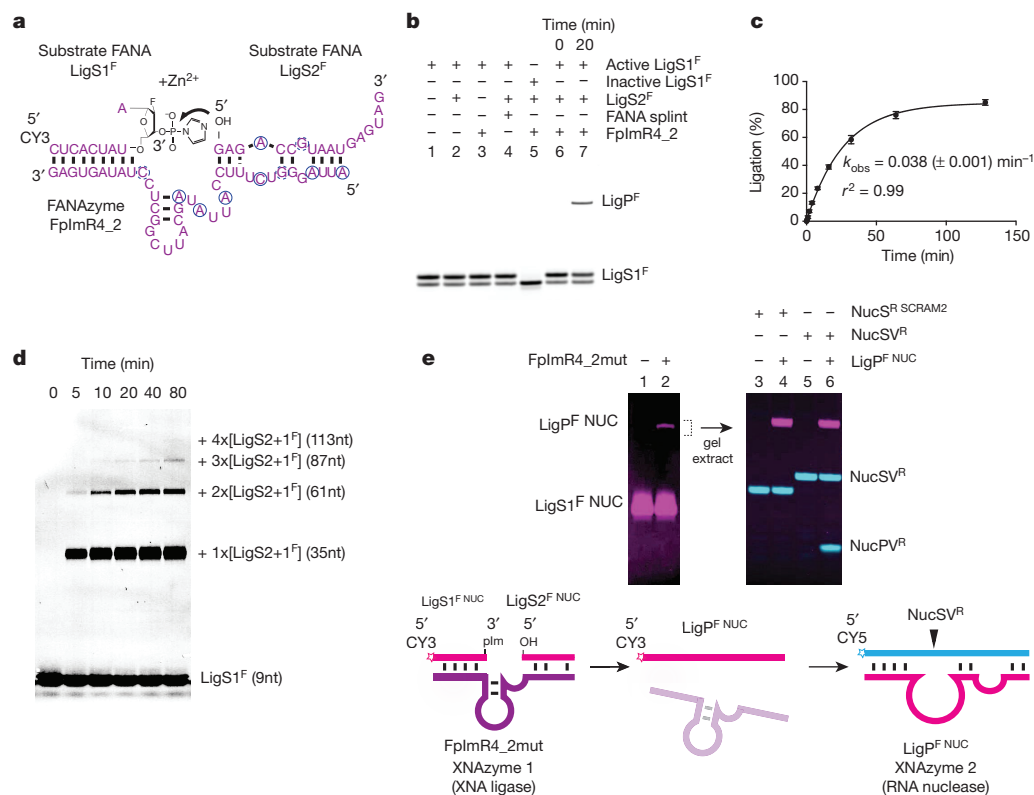


Figure 4 | XNA-XNA ligase XNAzyme (FANA) demonstrates catalysis without natural nucleic acids. **a**, Secondary structure (determined by DMS mapping, Extended Data Fig. 5) of chemically synthesized FANAzyme FpImR4_2, which ligates FANA LigS1^F, activated with 3' phosphorylimidazole (plm), to LigS2^F in *trans*. **b**, Urea-PAGE gel showing no product with: substrate LigS1^F alone (lane 1), no XNAzyme (lane 2), no LigS2^F (lane 3), splint (lane 4), or LigS1^F lacking 3' plm (lane 5); product formation is dependent on LigS2^F, activated LigS1^F and XNAzyme (lanes 6 and 7). **c**, Pre-steady state trimolecular reaction rate (k_{obs}) at 35 °C ($n = 3$; error

bars, s.d.). **d**, Urea-PAGE gel showing FpImR4_2-catalysed oligomerization of XNA (FANA) substrates. Substrate LigS2+1^F is a 3' plm-activated substrate containing the sequences of both LigS1^F and LigS2^F above. **e**, Urea-PAGE gels and schematic diagram showing XNAzyme-catalysed assembly of an active XNAzyme. A variant XNA ligase (FpImR4_2mut) catalyses ligation (lane 2) of FANA substrates LigS1^F NUC and LigS2^F NUC. The product (LigP^F NUC) is a variant of XNAzyme FR17_6 min (Fig. 2), which cleaves RNA substrate NucSV^R (lanes 5 and 6), but not scrambled RNA (NucS^R SCRAM2) (lanes 3 and 4).

sequence spaces for exploration, but their phenotypic richness remains to be determined. We have shown the discovery of catalysts (RNA endonucleases) in four such XNA sequence spaces (ANA, FANA, HNA, CeNA) and the elaboration of three different catalytic activities (RNA endonuclease, RNA ligase and XNA ligase) in one (FANA). These results indicate that properties such as catalysis (as well as heredity and evolution) are generalizable to a range of nucleic acid scaffolds and are likely to be emergent properties of many synthetic genetic polymers. This argues against a strong functional imperative for the chemistry of life's genetic systems.

Limitations in current XNA technology (for example, XNA-specific sequence biases, lower fidelity and sensitivity) contribute to library under-sampling, genetic drift and reduced selection stringency, complicating comparisons of phenotypic richness of the respective XNAs with DNA and RNA sequence spaces. Nevertheless, we note that the FANA framework, with similar hybridization energetics and conformational analogy to DNA¹⁹, yielded the most active XNAzymes, while catalysts in other XNAs, which exhibit reduced (ANA)¹³ or enhanced (HNA and CeNA) duplex stability, as well as divergent helical conformations and dynamics^{7,25}, showed slower rates. Substrate binding that is too weak or too strong, or conformational dynamics that are either too rapid or too slow, will reduce catalytic power by slowing conformational transitions required for catalysis and stabilizing inactive XNAzyme conformers. The evolutionary landscape of structurally more divergent XNAs may extend beyond the narrow parameters of DNA and RNA, suggesting that, for example, more effective HNA- or CeNAzymes might be discovered under non-physiological conditions. More work will be needed to resolve the question of whether life's reliance on RNA and DNA reflects a potential functional privilege of the natural polymers over unnatural XNAs in an ambient terrestrial environment⁹ or a predisposition of prebiotic chemistry.

Future advances in methodologies for the synthesis, replication and evolution of chemically ever more divergent genetic polymers should help to resolve these questions, providing a growing database of the molecular limits of chemical encoding and replication of information, while also yielding XNA catalysts (and ligands) that fully exploit their expanded range of physicochemical properties and biostability^{4,26–28} with potential applications ranging from medicine to nanotechnology.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 10 July; accepted 20 October 2014.

Published online 1 December 2014.

- Atkins, J. F., Gesteland, R. F. & Cech, T. R. (eds) *RNA Worlds* (Cold Spring Harbor Laboratory, 2012).
- Breaker, R. R. & Joyce, G. F. A DNA enzyme that cleaves RNA. *Chem. Biol.* **1**, 223–229 (1994).
- Eschenmoser, A. Chemical etiology of nucleic acid structure. *Science* **284**, 2118–2124 (1999).
- Pinheiro, V. B. *et al.* Synthetic genetic polymers capable of heredity and evolution. *Science* **336**, 341–344 (2012).
- Noronha, A. M. *et al.* Synthesis and biophysical properties of arabinonucleic acids (ANA): circular dichroic spectra, melting temperatures, and ribonuclease H susceptibility of ANA-RNA hybrid duplexes. *Biochemistry* **39**, 7050–7062 (2000).
- Wilds, C. J. 2'-Deoxy-2'-fluoro- β -D-arabinonucleosides and oligonucleotides (2'-F-ANA): synthesis and physicochemical studies. *Nucleic Acids Res.* **28**, 3625–3635 (2000).
- Herdewijn, P. Nucleic acids with a six-membered 'carbohydrate' mimic in the backbone. *Chem. Biodivers.* **7**, 1–59 (2010).

- Pinheiro, V. B. & Holliger, P. The XNA world: progress towards replication and evolution of synthetic genetic polymers. *Curr. Opin. Chem. Biol.* **16**, 245–252 (2012).
- Benner, S. A., Ricardo, A. & Carrigan, M. A. Is there a common chemical model for life in the universe? *Curr. Opin. Chem. Biol.* **8**, 672–689 (2004).
- Valadkhan, S. & Manley, J. L. Splicing-related catalysis by protein-free snRNAs. *Nature* **413**, 701–707 (2001).
- Nissen, P., Hansen, J., Ban, N., Moore, P. B. & Steitz, T. A. The structural basis of ribosome activity in peptide bond synthesis. *Science* **289**, 920–930 (2000).
- Joyce, G. F. Directed evolution of nucleic acid enzymes. *Annu. Rev. Biochem.* **73**, 791–836 (2004).
- Martín-Pintado, N. *et al.* The solution structure of double helical arabinonucleic acids (ANA and 2'-F-ANA): effect of arabinoses in duplex-hairpin interconversion. *Nucleic Acids Res.* **40**, 9329–9339 (2012).
- Lilley, D. M. J. Mechanisms of RNA catalysis. *Phil. Trans. R. Soc. Lond. B* **366**, 2910–2917 (2011).
- Wilkinson, K. A., Merino, E. J. & Weeks, K. M. Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. *Nature Protocols* **1**, 1610–1616 (2006).
- Homan, P. J. *et al.* Single-molecule correlated chemical probing of RNA. *Proc. Natl Acad. Sci. USA* **111**, 13858–13863 (2014).
- Santoro, S. W. & Joyce, G. F. A general purpose RNA-cleaving DNA enzyme. *Proc. Natl Acad. Sci. USA* **94**, 4262–4266 (1997).
- Santoro, S. W. & Joyce, G. F. Mechanism and utility of an RNA-cleaving DNA enzyme. *Biochemistry* **37**, 13330–13342 (1998).
- Minasov, G., Teplova, M., Nielsen, P., Wengel, J. & Egli, M. Structural basis of cleavage by RNase H of hybrids of arabinonucleic acids and RNA. *Biochemistry* **39**, 3525–3532 (2000).
- Bartel, D. P. & Szostak, J. W. Isolation of new ribozymes from a large pool of random sequences [see comment]. *Science* **261**, 1411–1418 (1993).
- Paul, N., Springsteen, G. & Joyce, G. F. Conversion of a ribozyme to a deoxyribozyme through *in vitro* evolution. *Chem. Biol.* **13**, 329–338 (2006).
- Eklund, E. H., Szostak, J. W. & Bartel, D. P. Structurally complex and highly active RNA ligases derived from random RNA sequences. *Science* **269**, 364–370 (1995).
- Rogers, J. & Joyce, G. F. The effect of cytidine on the structure and function of an RNA ligase ribozyme. *RNA* **7**, 395–404 (2001).
- Cuenoud, B. & Szostak, J. W. A DNA metalloenzyme with DNA ligase activity. *Nature* **375**, 611–614 (1995).
- Lescrier, E. *et al.* Solution structure of an HNA-RNA hybrid. *Chem. Biol.* **7**, 719–731 (2000).
- Dowler, T., Bergeron, D. & Tedeschi, A. L. Improvements in siRNA properties mediated by 2'-deoxy-2'-fluoro- β -D-arabinonucleic acid (FANA). *Nucleic Acids Res.* **34**, 1669–1675 (2006).
- Hendrix, C. *et al.* 1',5'-Anhydrohexitol oligonucleotides: hybridisation and strand displacement with oligoribonucleotides, interaction with RNase H and HIV reverse transcriptase. *Chemistry* **3**, 1513–1520 (1997).
- Nauwelaerts, K., Fisher, M. & Froeyen, M. Structural characterization and biological evaluation of small interfering RNAs containing cyclohexenyl nucleosides. *J. Am. Chem. Soc.* **129**, 9340–9348 (2007).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported by the Medical Research Council (MRC) programme grant U105178804 (P. Holliger, A.I.T., V.B.P., A.S.M., S.P.-C., C.C.) and by grants from the European Science Foundation (ESF) and the Biotechnology and Biological Sciences Research Council (BBSRC) UK (09-EuroSYNBIOP-013) (P.H., A.I.T.), the European Union Framework (FP7/2007-2013 (P. Herdewijn)), the European Research Council (ERC-2012 ADG_20120216/320683 (P. Herdewijn)), the US National Science Foundation (MCB-1121024 (K.M.W.)) and by an NSF Graduate Research Fellowship (DGE-1144081 (M.J.S.)).

Author Contributions A.I.T. and P. Holliger conceived and designed the experiments. A.I.T. performed XNAzyme selections and characterized XNAzymes with V.B.P., A.S.M., S.P.-C., C.C. V.B.P. generated the improved HNA synthetase. M.J.S. and K.M.W. performed and analysed SHAPE and DMS mapping experiments. P. Herdewijn synthesized hNTPs, ceNTPs and aGTP. All authors analysed data and co-wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to P. Holliger (ph1@mrclmb.cam.ac.uk).

METHODS

Nucleotides and oligonucleotides. Triphosphates of HNA (hNTPs), CeNA (ceNTPs) and ANA aGTP were synthesized and analysed as described previously⁴. Triphosphates of ANA (aATP, aCTP, aUTP) were obtained from TriLink BioTechnologies, FANA (faNTPs) from Metkinen Chemistry (Finland) and DNA (Illustra dNTPs) from GE Life Sciences (USA). Oligonucleotides were synthesized by Integrated DNA technologies (Belgium) or Sigma Aldrich (USA), unless stated otherwise. Triphosphorylated RNA (LigS2^R) was obtained from Trilink BioTechnologies (USA). Mock_LigP^R [2'-5'] and Mock_LigP^R [3'-5'] RNA standards were obtained from ChemGenes (USA). All oligonucleotides were purified by denaturing urea-PAGE and ethanol-precipitated from filtrates of freeze-thawed gel mash as described previously⁴.

Synthesis of XNAs. FANA and chimaeric DNA-FANA oligonucleotides were prepared either enzymatically (see below) or by solid-phase chemical synthesis using a Mermade4 instrument (BioAutomation, USA) with 1 µmol scale 3' phosphate (Synbase 1000, Link Technologies, UK) or universal (UnySupport 1000, Glen Research, USA) CPG supports. Phosphoramidites of DNA and all synthesis reagents were obtained from Link Technologies (UK), unless stated otherwise. The solid-phase synthesis method was adapted from Delevey *et al.*²⁹. Phosphoramidites of 2' fluoroarabinonucleosides (FANA), cyanine 3 fluorophore (CY3) or biotin-triethyleneglycol (BiotinTEG) were obtained from Glen Research (USA) and prepared as 0.15 M solutions in anhydrous acetonitrile (ACN) (Sigma Aldrich, USA), those of DNA were prepared as 0.1 M solutions. Phosphoramidites were activated with 0.3 M BTT (5-benzylthio-1H-tetrazole in ACN), deblocking was performed with 3% trichloroacetic acid in dichloromethane, capping of failure sequences was performed with pyridine acetic anhydride and 10% methylimidazole in tetrahydrofuran, and oxidation was performed with 0.02 M iodine oxidiser (Prologo series, Sigma Aldrich, USA). Coupling times were 600 s for all phosphoramidites, with the exception of CY3, BiotinTEG and the FANA guanosine phosphoramidite, which were allowed to couple for 900 s. Deprotection and cleavage from CPG support was achieved by incubation in 3:1 NH₄OH:EtOH for 48 h at room temperature, then dried by speedvac. PAGE purified chemically synthesized FANA substrates and XNAzymes were analysed by mass spectrometry (Extended Data Fig. 10). For triphosphate addition to synthesized FANA (and DNA), the method described by Zlatev *et al.*³⁰ was followed before deprotection and cleavage from the solid support.

All other XNAs were prepared enzymatically using polymerase mutants as described previously⁴; polymerases D4K for ANA and FANA, 6G12 for CeNA and 6G12 I521L (see below) for 1,5 anhydrohexitol nucleic acid (HNA), with the addition of 4% ET-SSB (NEB, USA). For preparation of all-XNA strands using polymerases (for example, for *trans* XNAzyme reactions), either the appropriate FANA primer was used, or an RNA primer was used to synthesis an RNA-XNA chimaeric strand, which was then incubated in 0.8 M NaOH at 65 °C for 1 h to completely hydrolyse the RNA portion. All XNA oligonucleotides were purified by denaturing urea-PAGE.

Preparation of single-stranded DNA, RNA and XNA. Biotinylated oligos were captured using Dynabeads MyOne Streptavidin C1 beads (Invitrogen/Life Technologies, USA) in BWBS (10 mM Tris-HCl pH 7.4, 1 M NaCl, 0.1% v/v Tween20, 1 mM EDTA) for 1–2 h at room temperature or overnight at 4 °C. Denaturation/elution of unbiotinylated strands was achieved by three washes in BWBS followed by rapid (<1 min) incubation in 0.1 M NaOH at room temperature. Where eluted strand was being prepared, NaOH supernatant was immediately neutralized in 1 M Tris pH 7.4. Elution of biotinylated oligos from beads was achieved by three washes in H₂O, then incubation in either H₂O for 2 × 2 min at 80 °C, or PAGE loading buffer (95% formamide, 10 mM EDTA, 0.05% bromophenol blue) for 2 × 2 min at 95 °C.

Preparation of 6G12 I521L polymerase. We introduced the I521L mutation to the 6G12 backbone by iPCR using primers RT520fo and RT521ba. PCR was carried out using Expand High Fidelity polymerase (Roche Diagnostics GmbH, Germany) as an initial incubation of 2 min at 95 °C followed by 25 × of (30 s 95 °C, 30 s 50 °C, 18 min 68 °C) followed by a final extension of 10 min 68 °C. Amplified DNA was purified (QIAquick PCR purification kit, Qiagen GmbH, Germany) according to the manufacturer's recommendations and restricted with BsaI and DpnI (New England Biolabs Inc., Massachusetts, USA). Reactions were again purified (QIAquick PCR purification kit) and ligated with T4 DNA ligase (NEB, USA). Ligated plasmids were transformed into *E. coli* NEB 10-β cells (NEB, USA), and isolated transformants were checked by DNA sequencing (Source Biosciences, UK).

A transformant with the correct sequence was expressed and purified as previously described⁴ and used to determine the impact of the additional mutation on the fidelity and processivity of 6G12. The resulting 6G12 I521L polymerase had a different divalent cation optimum and could synthesize HNA in the absence of Mn²⁺ ions, in reactions carried out with 3 mM Mg²⁺. 6G12 I521L was more processive than 6G12 alone and could synthesize HNA at higher fidelities (aggregate DNA → HNA → DNA fidelity: 3.0 × 10⁻³ — experiment carried out as described previously⁴).

Preparation of FANA phosphorylimidazolid oligonucleotides. Preparation of phosphorylimidazolid oligonucleotides was adapted from a method used by Orgel and others³¹. 3' phosphorylated FANA or DNA was prepared by solid-phase chemical synthesis (see above) and re-suspended to 100 µM in 0.5 M imidazole (pH 6.0). 50 µl

oligo/imidazole solution was added to 6.5 µmol solid 1-ethyl-3-[3-dimethylamino-propyl]carbodiimide hydrochloride (EDC) (Pierce Biotechnology/Thermo, USA) and incubated at room temperature for 2 h. Oligos were desalted using Amicon 3,000 MW cut-off spin filters (Merck Millipore, USA). Purification of (and analysis of reactions involving) all phosphorylimidazolid oligos were performed using Tris-free urea-PAGE gels run using 10 mM NaOH, pH adjusted to 8.5 with boric acid³². FANA phosphorylimidazolid oligos were analysed by mass spectrometry, phosphatase protection and urea-PAGE mobility (Extended Data Fig. 10).

XNAzyme selections. General schemes for selections are shown in Extended Data Figs 1, 6 and 8. Purified single-stranded libraries and substrates were annealed by incubation at 80 °C for 60 s, then allowed to cool to room temperature over 5 min, except for phosphorylimidazolid oligos, which were not annealed. Substrates and enzymes were incubated separately in reaction buffer for 5 min at reaction temperature, then mixed to start reactions. For selection of RNA ligase and endonuclease XNAzymes, reactions were performed at 17 °C in 30 mM HEPES (pH 8.5), 150 mM KCl, 25 mM MgCl₂ and 0.5 U µl⁻¹ RNasein RNase inhibitor (Promega, USA). For selection of XNA (FANA) ligase XNAzymes, reactions were performed at 35 °C in 30 mM HEPES (pH 7.2), 150 mM KCl, 25 mM MgCl₂ and 1 mM ZnCl₂. In general, ~1 nmol of starting library was prepared and reacted at 10 µM with equimolar substrate for 5 days. For rounds 2–17, 10–50 pmol XNA pools were prepared and reacted at 1 µM over steadily decreasing reaction times, settling on 30 min in rounds 15–17. In RNA endonuclease XNAzyme selections, the three libraries for each XNA (fully degenerate N40 library, and the '8-17' and '10-23' patterned libraries) were synthesized separately, but pooled after round 5.

All XNA reverse transcriptions (using polymerase RTI521L) were performed as described previously for HNA aptamer selections⁴, but without a polyA tailing step and using 0.2 µM RT primer Tag4test7_2Me (or a version with 5' BiotinTEG), which contains 2'-O-methyl RNA modifications to improve annealing. First-stand cDNA was amplified by a two-step nested PCR strategy (see Extended Data Figs 1, 6 and 8). The first 'out-nested' RT + PCRs used 0.5 µM primers and a mixture of OneTaq Hot Start (NEB, USA) (itself a mix of *Taq* and Deep Vent_R) and 0.15 U µl⁻¹ Thermoscript (Invitrogen/Life Technologies, USA) polymerases, which is able to transcribe 2'-O-methyl RNA, in 20 mM Tris-HCl (pH 8.9 at 25 °C), 22 mM NH₄Cl, 22 mM KCl, 0.06% IGEPAL CA-630, 0.05% Tween20, 4 mM MgCl₂ and 200 µM dNTPs. Cycling conditions were 80 °C for 30 s, 52 °C for 30 s, 72 °C for 15 min, 94 °C for 1 min, 20–35 × [94 °C for 30 s, 54 °C for 30 s, 72 °C for 30 s], 72 °C for 2 min. Following the first PCR, primers were digested using ExoSAP (Ambion/Life Technologies, USA), which was then heat inactivated, according to the manufacturer's instructions. Second step ('in-nest') PCRs used using 1 µl of unpurified out-nest PCR product as template in a 50 µl reaction using OneTaq Hot Start master mix (NEB, USA) and cycling conditions 94 °C for 1 min, 10–20 × [94 °C for 30 s, 54 °C for 30 s, 72 °C for 30 s], 72 °C for 2 min. Reactions were analysed by electrophoresis on 4% NGQT-1000 agarose (Thistle Scientific, UK) gels containing GelStar stain (Lonza, Switzerland). Bands of appropriate size were purified using a gel extraction kit (Qiagen, Netherlands) as per manufacturer's instructions. Purified DNA was used as the polyclonal template for either sequencing library PCR (see below) or large scale preparative PCR (2 ml) for generation of DNA templates for XNA synthesis. Prep PCR were performed with 1 µM primers using 0.05 U µl⁻¹ SUPER Taq in 1X buffer (HT Biotechnology, UK) with 0.125 µM dNTPs. Cycling conditions were the same as the second step PCR above. Single-stranded DNA templates were isolated using streptavidin beads (see above) and ethanol-precipitated before further use.

XNAzyme reactions. Purified XNAzymes and substrates were annealed as described above and reacted under selection conditions unless stated otherwise, in DNA- or protein- (for 3' pIm reactions) LoBind tubes (Eppendorf, Germany). In pH titration experiments, buffer was substituted for 50 mM EPPS (pH 6.5–8.75), CHES (pH 9.0–10.0) or CAPS (pH 10.25–11.0). For determination of pseudo first-order reaction rate (k_{obs}) under single-turnover pre-steady-state (K_m/k_{cat}) conditions, a fivefold excess of enzyme (5 µM) was incubated in *trans* with either fluorophore-labelled 1 µM NucS (nuclease substrate), or fluorophore-labelled 1 µM LigS1 (ligase substrate 1) and 5 µM LigS2 (ligase substrate 2). RNA endonuclease and ligase reactions were performed in 30 mM EPPS (pH 8.5), 150 mM KCl, 50 mM MgCl₂ at 25 °C. XNA ligase reactions were performed in 30 mM HEPES (pH 7.5), 150 mM KCl, 50 mM MgCl₂, 1 mM ZnCl₂ at 35 °C. Reactions were stopped at different time points by addition of 95% formamide, 20 mM EDTA and cooling on dry ice. Reactions were separated by urea-PAGE and fluorophores visualized using a Typhoon Trio imager (GE Life Sciences, UK). The fraction of reaction product to substrate was quantified using ImageQuant TL software (GE Life Sciences, UK) and mean data from three independent reactions (except for CeNAzyme CeR16_3, for which only two data sets were collected) were fitted to equation (1) using Prism 6.0b (GraphPad, USA):

$$P(t) = P_{\infty} (1 - e^{-k_{obs}t}) \quad (1)$$

where $P(t)$ is the percentage of cleaved or ligated RNA or XNA (FANA) at time t , P_{∞} is the apparent reaction end point and k_{obs} is the observed rate constant. For magnesium titration experiments, data were fitted to equation (2):

$$P(t) = P_{\infty} \times \frac{[\text{Mg}^{2+}]}{K_m + [\text{Mg}^{2+}]} \quad (2)$$

where $P(t)$ is the percentage of cleaved or ligated RNA or XNA (FANA) at time t , P_{∞} is the apparent reaction end point and K_m is the apparent Michaelis constant.

For FR17_6min multiple turnover catalysis, 1 μM NucSR_min(AG) was reacted with 10 nM FR17_6min at 25 °C in 30 mM HEPES (pH 8.5), 150 mM KCl, 50 mM MgCl_2 . DNAzyme 8-17 synthesized as XNAs (5 μM) was reacted with substrate NucSR_min(AG) (1 μM) for 1 h at 37 °C in 30 mM HEPES (pH 8.5), 150 mM NaCl, 50 mM MgCl_2 . For characterization of bivalent metal ion requirements of selected XNAzymes, bimolecular (FR17_6min) or trimolecular (F2R17_1min and FpImR4_2) reactions were performed under conditions used to determine k_{obs} with MgCl_2 and/or ZnCl_2 substituted by chlorides of the metals of the Irving–Williams series (100 μM –50 mM).

Deep sequencing. Amplified polyclonal cDNA from XNA selections was prepared for deep sequencing by the Illumina Miseq method by appending the bridge-amplification sequences by PCR. Sequencing library generating PCR reactions were performed with OneTaq Hot Start master mix (NEB, USA) with 10 ng per 50 μl gel-purified polyclonal template DNA (see above), 0.1 μM primers (P5_P2 and P3_Test7-2) and cycling conditions 94 °C for 1 min, 10 \times [94 °C for 30 s, 56 °C for 30 s, 72 °C for 30 s], 72 °C for 2 min. Sequencing library DNA was purified using a PCR purification kit (Qiagen, Netherlands), then a 12 pM sample of pooled libraries plus 20% PhiX control (Illumina, UK) was denatured and sequenced (single-end read, 75 cycles) using a MiSeq reagent kit and instrument (Illumina, UK) according to manufacturer's instructions. Libraries were barcoded using variants of the P5_P2 primer containing 6 nt sequences from the NEXTFlex series (Illumina, UK). Data were analysed using the Galaxy server^{33–35} and sequences ordered by abundance.

Analysis of oligonucleotide phosphorylation. The presence or absence of 3' or 5' phosphates on RNA cleavage products, and the protection of 3' phosphates on FANA ligase substrates by formation of phosphorylimidazolides, was assayed by urea–PAGE gel shift following incubation in rAPid alkaline phosphatase (Roche, Switzerland) or T4 polynucleotide kinase (NEB, USA) in manufacturer's buffers for 30 min at 37 °C. Hydrolysis of cyclic phosphates was achieved by incubation in 10 mM glycine pH 2.5 for 10 min at room temperature. Partial alkaline hydrolysis of RNA substrates (denoted by ^-OH) was achieved by incubation at 90 °C in 50 mM sodium carbonate buffer (pH 9.2) for 10 min. Partial RNase T1 digestion was achieved by incubation at 55 °C in 0.1 U μl^{-1} RNase T1 (Invitrogen/Life Technologies, USA) in 30 mM sodium acetate (pH 5) for 10 min, then stopped in 7 M urea 1.5 mM EDTA.

Analysis of oligonucleotide mass by MALDI-ToF mass spectrometry. Oligo samples, 0.75 μl in water, were spotted onto a MALDI target followed by 0.75 μl of 3-hydroxypicolinic acid. Some oligo samples were vacuum dried, resuspended in 25 μl , 0.1 M TEAA (triethylammonium acetate) and desalted using a zip-tip C18 (Merck Millipore, USA). The zip-tip C18 was washed 3 \times 10 μl of 0.1 M TEAA and then 3 \times 10 μl water. Next, the oligo was eluted directly onto a MALDI target with 5 μl of 3-hydroxypicolinic acid. All mass spectrometric measurements were carried out in positive ion mode on an Ultraflex III TOF-TOF instrument (Bruker Daltonik, Bremen, Germany).

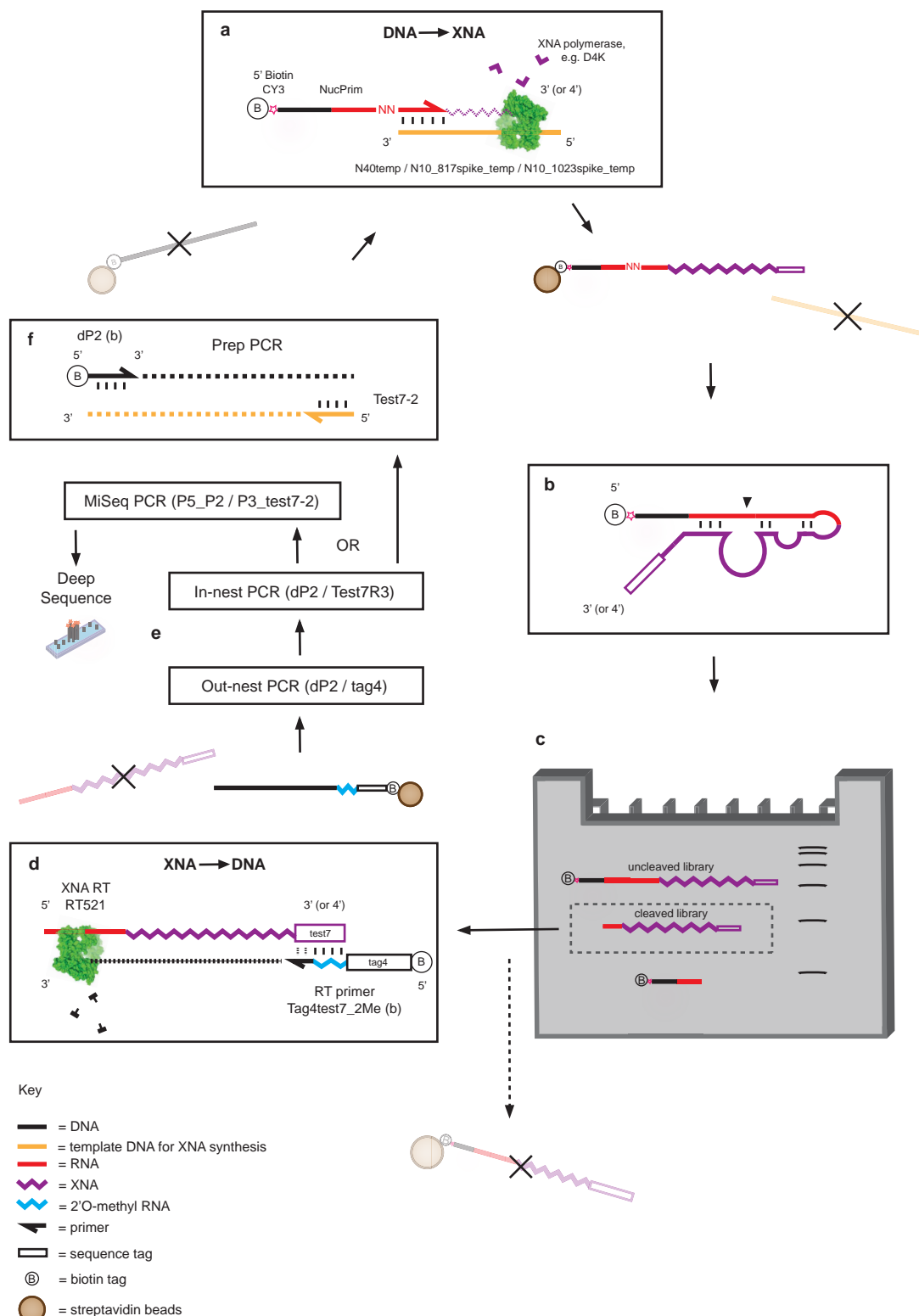
Analysis of oligonucleotide linkage isomers by SAX-HPLC. RNA and FANA ligation products were analysed by strong anion exchange chromatography (SAX-HPLC) using a Varian Prostar system (Agilent, USA) with a DNAPac PA200 column (Dionex/Thermo, USA) under conditions sufficient to resolve linkage regioisomers³⁶. 10 nM sodium phosphate buffer (pH 11.5), gradient 0.4 M to 1.4 M NaCl over 30 min, flow rate 1.5 ml min^{-1} . Fluorescence was detected using a 122 fluorometer (Gilson, USA) set to excitation 488 nm, emission 520 nm for carboxyfluorescein (6FAM)-labelled RNA, and emission 550 nm, excitation 570 nm for cyanine 3 (CY3)-labelled FANA.

Chemical probing of XNAzymes and secondary structure prediction. Selective 2'-hydroxyl acylation analysed by primer extension (SHAPE) structure probing experiments were performed on RNA endonuclease XNAzyme (FANA) FR17_6 using a chimaeric RNA–FANA construct, FR17_6wilc, prepared using primer SHAPE_Nucprim and template FR17_6wilc_temp. The construct contains sequences of NucSR^R (RNA) and FR17_6 (FANA) flanked by 5' and 3' structure cassette sequences from Wilkinson *et al.*¹⁵, with a 2'-O-methyl RNA modification at adenine 11. For the XNA ligase XNAzyme (FANA) FpImR4_2, an analogous construct, FpImR4_2wilc, was prepared by ligation of 1 μM modified FANA substrate LigS1wilc^F to an equimolar concentration of a version of FpImR4_2 with LigS2^F in *cis* (prepared by LigS2^F-primed FANA synthesis on template FpImR4wilc_temp) for 2 h at 35 °C in 30 mM HEPES (pH 7.2), 150 mM KCl, 25 mM MgCl_2 and 1 mM ZnCl_2 . FANA constructs (1 μM in 8 μl H₂O) were denatured at 80 °C for 1 min, incubated at room temperature for 5 min, treated

with 1 μl 10 \times SHAPE folding buffer (500 mM EPPS (pH 8.2), 1.5 M KCl, 250 mM MgCl_2), and allowed to fold at 17 °C for 20 min. After folding, FANA constructs were treated with 1-methyl-7-nitroisatoic anhydride (1 μl , 100 mM in neat DMSO) and incubated at 17 °C for 15 min. No-reagent control reactions were performed with 1 μl neat DMSO. Denaturing control reactions were performed as described previously³⁷. After modification, FANA constructs were purified with a G-50 spin column (GE Healthcare). cDNA was generated using SuperScript II reverse transcriptase (Life Technologies) under SHAPE-MaP conditions³⁷.

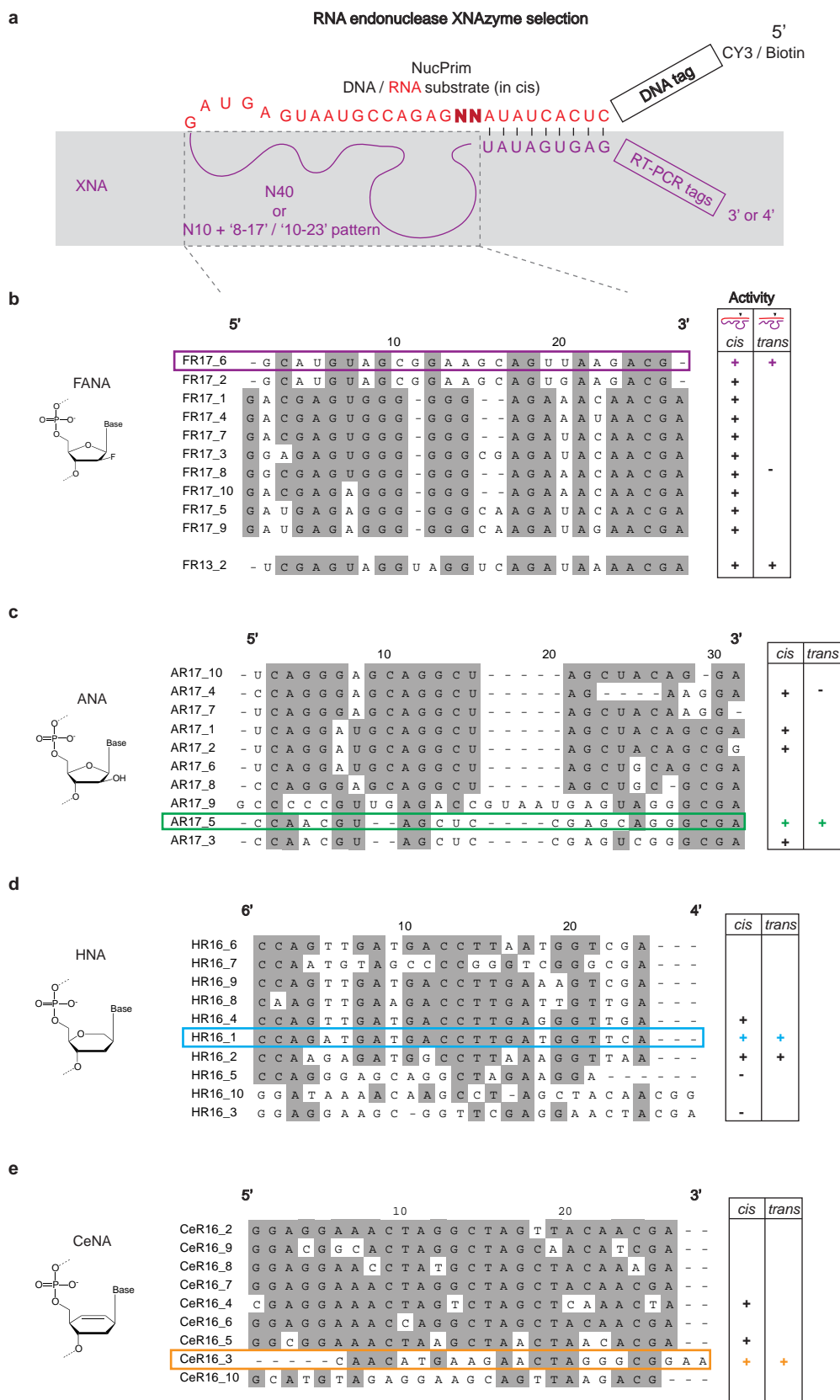
Dimethyl sulphate (DMS) modification was adapted from the RING-MaP approach¹⁶. FANA constructs (nuclease: 1 μM in 5 μl H₂O, ligase: 0.5 μM in 5 μl H₂O) were annealed as described above and treated with 4 μl 2.5 \times DMS folding buffer (750 mM cacodylate pH 7.0, 25 mM MgCl_2). Folded FANA constructs were treated with DMS (1 μl , 1.7 M in absolute ethanol), incubated at 17 °C for 6 min, quenched with 10 μl neat 2-mercaptoethanol, and purified with a G-50 spin column. No-reagent control reactions were performed with 1 μl absolute ethanol. cDNA was generated using RT521K polymerase as described previously⁴. Briefly, 5 pmol FANA construct and 10 pmol primer were denatured for 1 min at 95 °C, chilled on ice, and incubated with $\sim 2 \mu\text{g ml}^{-1}$ RT521K and 0.2 mM dNTPs in 1 \times ThermoPol buffer (New England Biolabs) at 65 °C for 4 h. cDNA from reverse transcription reactions was purified with G-50 spin columns. SHAPE- and DMS-MaP sequencing libraries were created using the targeted gene-specific approach³⁷, with minor changes: PCR 1 was performed for 23 cycles, 98 °C for 30 s, 23 \times [98 °C for 10 s, 68 °C for 30 s, 72 °C for 20 s], 72 °C for 2 min, and PCR 2 was performed for 7 cycles, using 1 μl of unpurified PCR 1 product as template in a 50 μl reaction. Purified libraries were pooled and sequenced with an Illumina MiSeq, generating data sets of 2 \times 150 paired-end reads. Sequencing reads were aligned to reference sequences and per-nucleotide mutation rates, excluding primer-binding sites, were calculated using the SHAPE-MaP analysis pipeline. SHAPE reactivities were calculated for RNA nucleotides³⁷; FANA nucleotides were excluded from SHAPE analysis. DMS reactivities for all nucleotides were calculated by subtracting the mutation rate of the no-reagent control from the mutation rate of DMS-modified FANAzymes at each position. SHAPE and DMS reactivity profiles were normalized by the '2%-8%' method³⁸. The FR17_6 FANAzyme secondary structure model is the only structure predicted using ShapeKnots³⁹, incorporating pseudo-free energy constraints derived from SHAPE reactivities. All other XNAzyme secondary structure models were predicted using ViennaRNA (version 2.1.6)⁴⁰ or mfold⁴¹. The FpImR4_2 structure was further manually curated using DMS reactivity data. Oligonucleotide sequences can be found in Supplementary Table 1.

29. Deleavey, G. F. *et al.* Synergistic effects between analogs of DNA and RNA improve the potency of siRNA-mediated gene silencing. *Nucleic Acids Res.* **38**, 4547–4557 (2010).
30. Zlatev, I., Manoharan, M., Vasseur, J.-J. & Morvan, F. Solid-phase chemical synthesis of 5'-triphosphate DNA, RNA, and chemically modified oligonucleotides. In *Current Protocols in Nucleic Acid Chemistry* <http://dx.doi.org/10.1002/0471142700.nc0128s50> (Wiley & Sons, 2001).
31. Chu, B. C., Wahl, G. M. & Orgel, L. E. Derivatization of unprotected polynucleotides. *Nucleic Acids Res.* **11**, 6513–6529 (1983).
32. Brody, J. R. & Kern, S. E. Sodium boric acid: a Tris-free, cooler conductive medium for DNA electrophoresis. *Biotechniques* **36**, 214–216 (2004).
33. Goecks, J., Nekrutenko, A., Taylor, J. & The Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* **11**, R86 (2010).
34. Blankenberg, D. *et al.* Galaxy: a web-based genome analysis tool for experimentalists. In *Current Protocols in Molecular Biology* Ch. 19, Unit 19.10.1–21 (2010).
35. Giardine, B. *et al.* Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.* **15**, 1451–1455 (2005).
36. Bowler, F. R. *et al.* Prebiotically plausible oligoribonucleotide ligation facilitated by chemoselective acetylation. *Nature Chem.* **5**, 383–389 (2013).
37. Siegfried, N. A., Busan, S., Rice, G. M., Nelson, J. A. E. & Weeks, K. M. RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nature Methods* **11**, 959–965 (2014).
38. Mortimer, S. A. & Weeks, K. M. A. Fast-acting reagent for accurate analysis of RNA secondary and tertiary structure by SHAPE chemistry. *J. Am. Chem. Soc.* **129**, 4144–4145 (2007).
39. Hajdin, C. E. *et al.* Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. *Proc. Natl Acad. Sci. USA* **110**, 5498–5503 (2013).
40. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **6**, 26 (2011).
41. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31**, 3406–3415 (2003).



Extended Data Figure 1 | Selection scheme for RNA endonuclease XNAzymes. **a**, XNA library preparation using DNA-dependent XNA polymerases, templated using N40temp, N10_817spike_temp, or N10_1023spike_temp DNA oligonucleotides (see Supplementary Table 1) primed by a biotinylated chimaeric DNA–RNA primer (NucPrim), which serves as substrate for RNA cleavage in *cis*. Libraries are captured by streptavidin beads, allowing cleavage and removal of DNA templates. **b**, Single-stranded libraries are annealed and incubated in reaction buffer (see Methods), successful XNAzymes cleave the biotinylated RNA substrate in *cis*. **c**, Size separation of reacted XNA pools using denaturing polyacrylamide

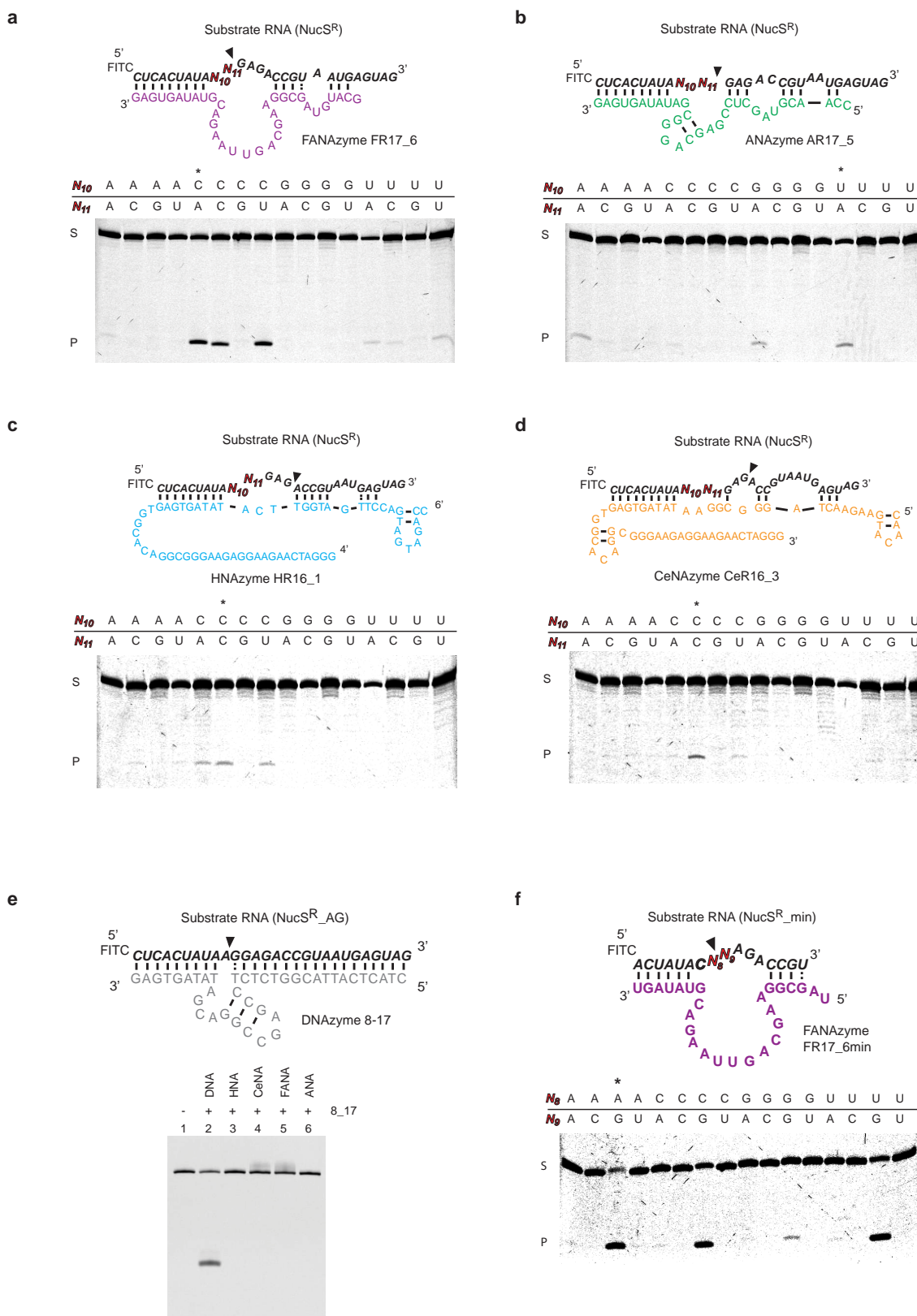
electrophoresis (urea–PAGE). Cleaved XNA pools are gel-extracted (indicated by dashed box) and incubated with streptavidin beads in order to capture and discard any uncleaved carry-over (indicated by dashed arrow). **d**, Reverse transcription of isolated, cleaved XNA pools using XNA-dependent DNA polymerase RT521L (that is, XNA → cDNA). **e**, Amplification of transcribed cDNA by successive PCR reactions, using the primers indicated (see Supplementary Table 1). **f**, PCR reaction generating templates for XNA synthesis for further rounds of selection. See Methods for details. Solid crosses indicate removal of denatured strands using streptavidin bead capture.



Extended Data Figure 2 | Sequences of RNA endonuclease XNAzymes.

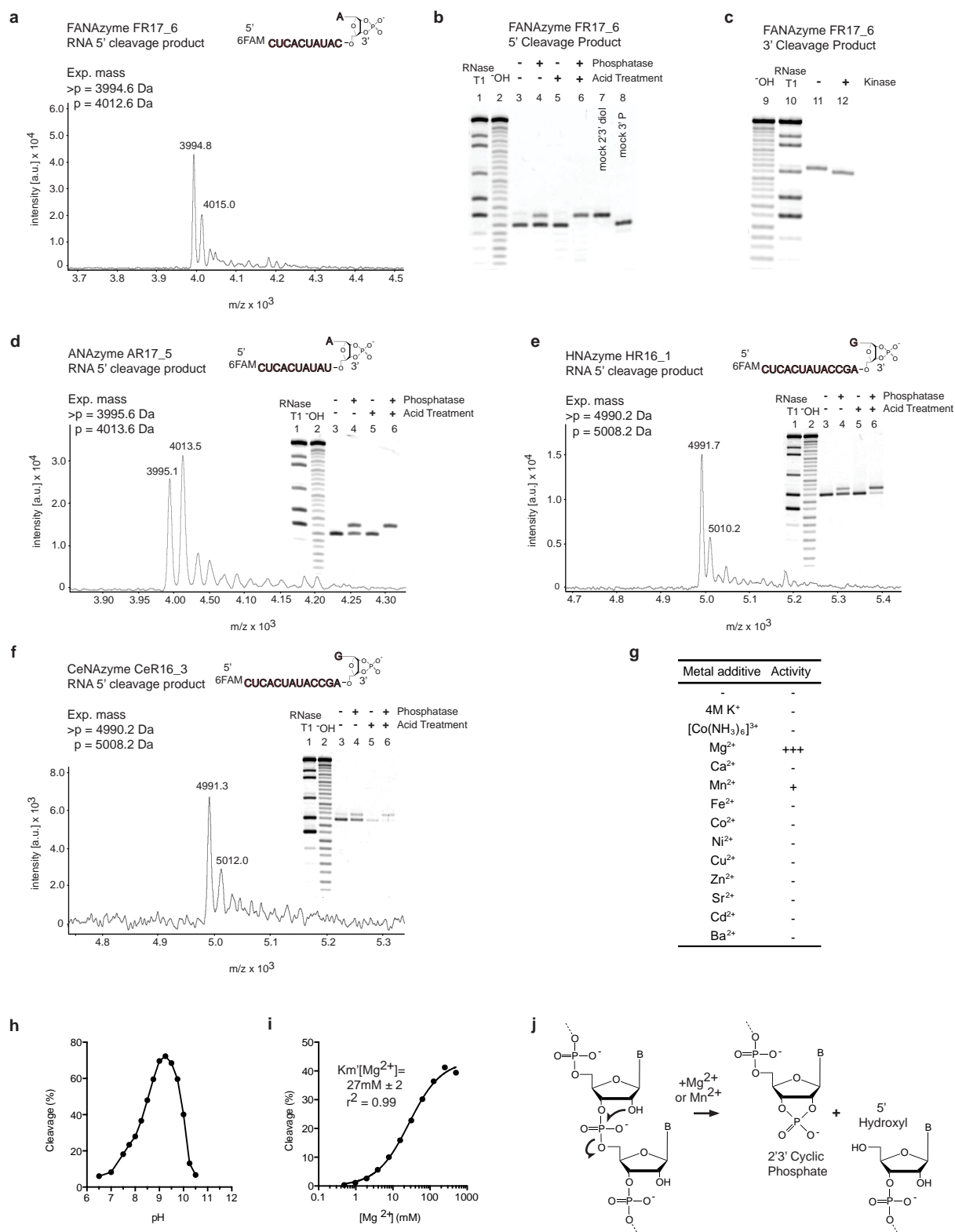
a, Schematic diagram showing DNA–RNA–(red)–XNA(purple) chimaeric library setup for selection of in *cis* RNA-cleaving XNAzymes. The sequences of the XNA region under selection (dashed box) of the most abundant clones revealed by deep sequencing are shown for selections using **b**, FANA, **c**, ANA,

d, HNA, and **e**, CeNA. The top 10 sequences, or representatives of sequence families, were screened by Urea–PAGE gel shift for activity in *cis* (unimolecular reaction, as selected) and in *trans* (bimolecular reaction). Sequences chosen for further characterization are indicated by coloured boxes.



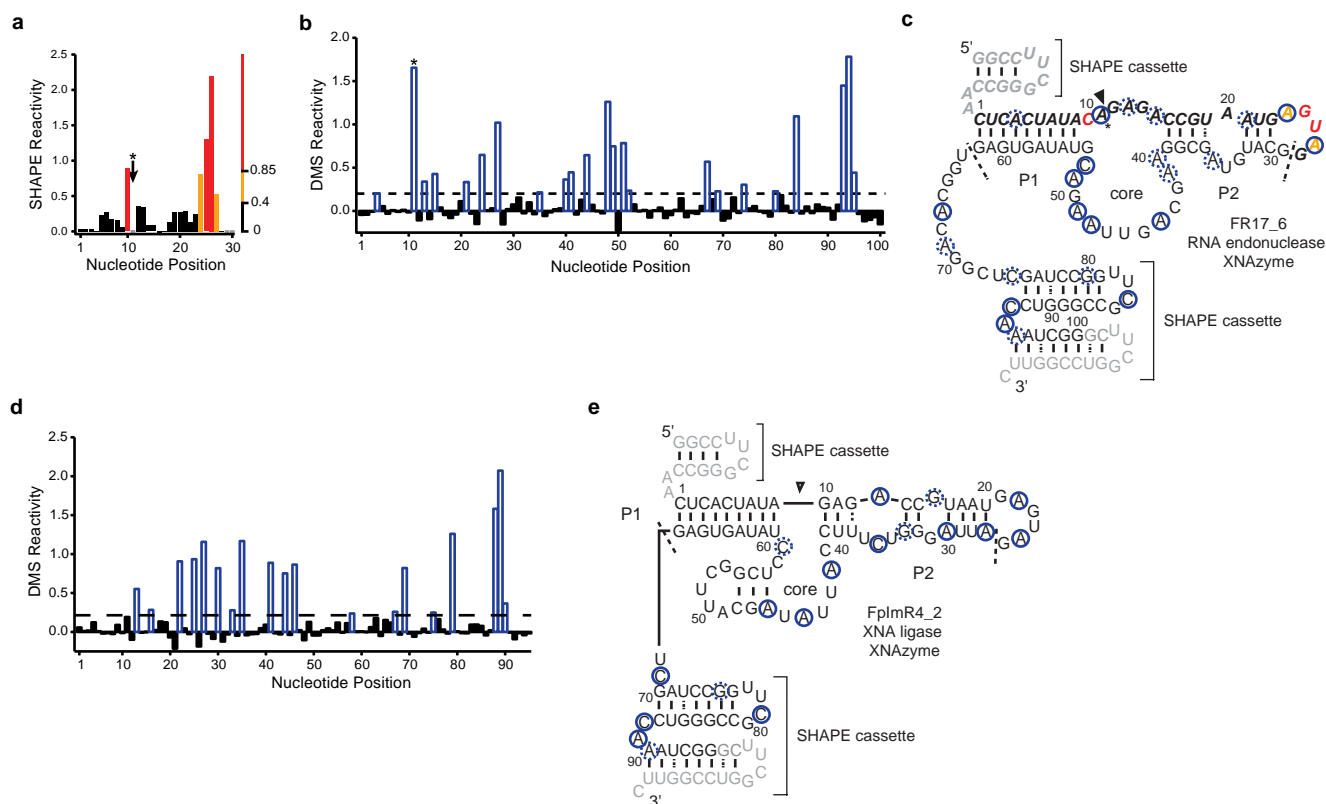
Extended Data Figure 3 | Sequence dependence of RNA endonuclease XNAzyme cleavage. XNAzymes were selected with degeneracy in the RNA substrate (see Extended Data Fig. 2a). The sequence requirements at these positions (upstream of the cleavage sites shown by a black inverted triangle) in the RNA substrate (N₁₀ and N₁₁ shown in red) were determined by urea-PAGE gel shift using all 16 variants of the substrate NucS^R with each XNAzyme in *trans*: **a**, FR17_6 (FANA), **b**, AR17_5 (ANA), **c**, HR16_1 (HNA), **d**, CeR16_3 (CeNA). **e**, RNA substrate NucS^R AG (lane 1) was reacted in *trans* with RNA

endonuclease DNAzyme 8-17¹⁷ synthesized as DNA (lane 2), HNA (lane 3), CeNA (lane 4), FANA (lane 5) or ANA (lane 6). Activity of 8-17 is lost upon conversion to the XNAs described. **f**, Sequence requirements for RNA cleavage by FANAZyme FR17_6 min proximal to the cleavage site (black inverted triangle), positions 8 and 9 (highlighted in red) of minimized RNA substrate NucS^R_min. Substrate sequences used for further characterization of each XNAzyme are indicated by an asterisk.



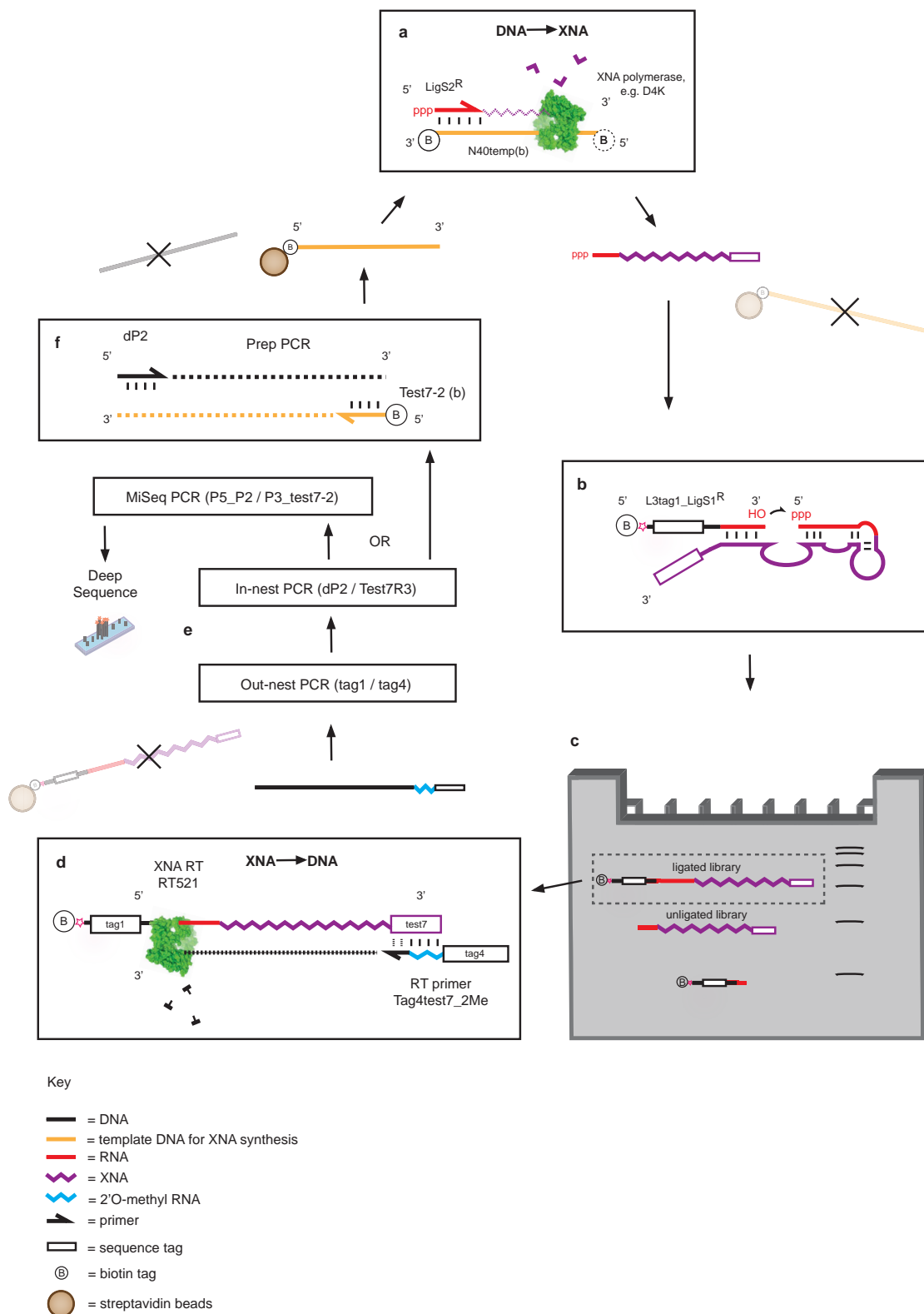
Extended Data Figure 4 | Analysis of RNA endonuclease XNAzyme cleavage products. **a**, 5' cleavage product of FANzyme FR17_6 reaction shows expected mass for a 2',3' cyclic phosphate (>p) using matrix-assisted laser desorption/ionization-time of flight mass spectrometry (MALDI-ToF). **b**, Hydrolysis of 5' FR17_6 cleavage product >p in low pH and dephosphorylation with calf intestinal phosphatase (removes 2'p or 3'p, but not >p). **c**, Phosphorylation of 3' FR17_6 cleavage product with T4

polynucleotide kinase (adds 5'p). Mass spectra and dephosphorylation assays of 5' cleavage products of **d**, ANAzyme AR17_5, **e**, HNAzyme HR16_1 and **f**, CeNAzyme CeR16_3 reveal all RNA endonuclease XNAzymes yield products with 2',3' cyclic phosphates. (RNase T1 and (-OH) indicate partial hydrolysis reactions of the RNA substrates used. **g**, Bivalent metal ion requirements and titration of, **h**, pH or **i**, MgCl₂, of FANzyme FR17_6min reaction with NucS^R_min. **j**, Reaction catalysed by RNA endonuclease XNAzymes.



Extended Data Figure 5 | Chemical probing of XNAzyme secondary structures. **a**, Chemical probing using selective 2'-hydroxyl acylation analysed by primer extension (SHAPE)(RNA)¹⁵ or, **b**, **d**, dimethyl sulphate (DMS)¹⁶ (RNA and FANA) footprinting, used to inform secondary structure predictions of **c**, RNA endonuclease FANAzyme FR17_6 or **e**, FANA ligase FANAzyme FpImR4_2, embedded in structural cassettes¹⁵, with RNA substrate (inactivated by a 2'-O-methyl modification at position 11, indicated by an asterisk in all panels) or FANA product in *cis*. SHAPE analyses RNA 2'-OH. Black, orange and red solid bars and bases indicate low, moderate, and high SHAPE reactivity,

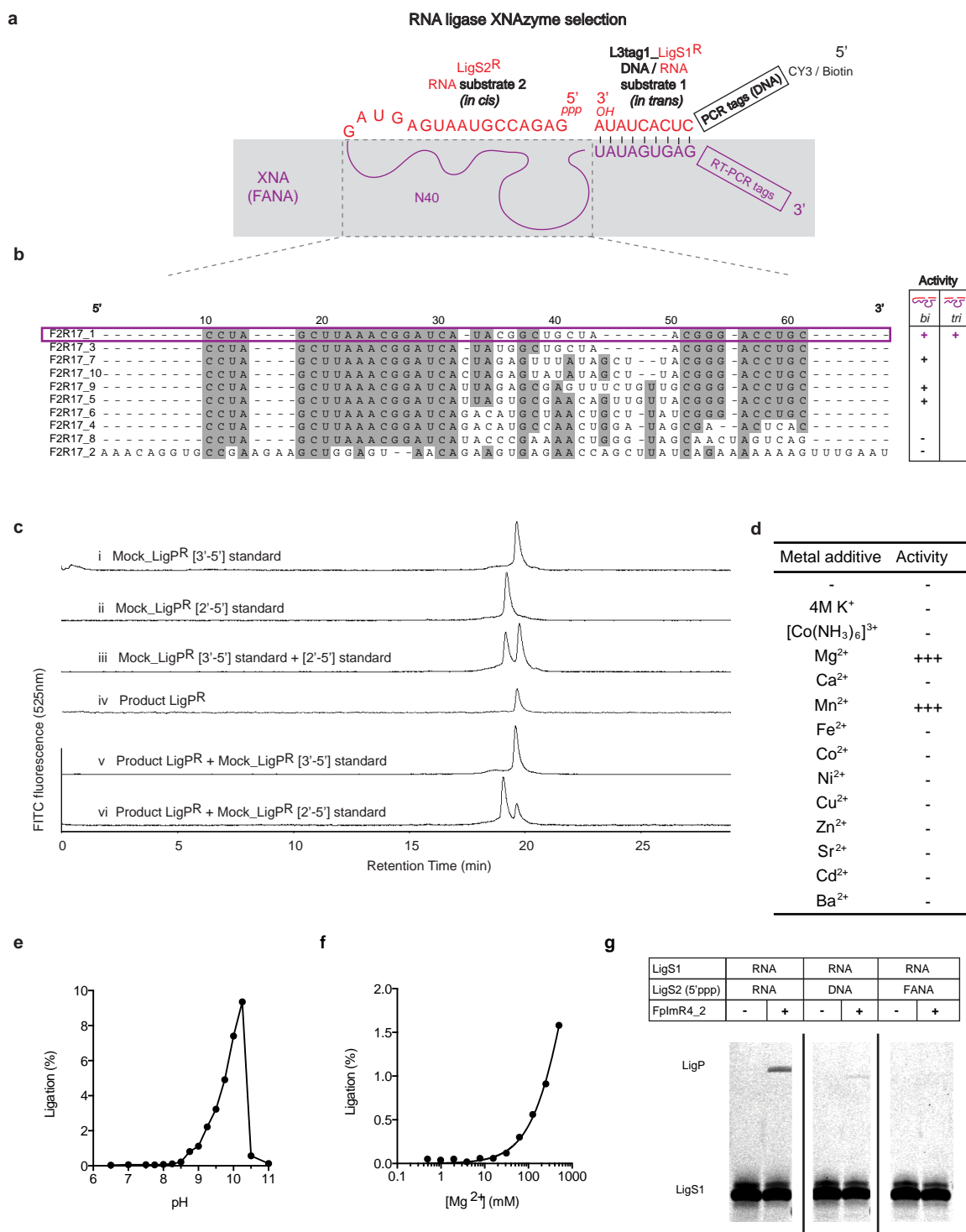
respectively, for positions 1-28 in FR17_6 construct, corresponding to RNA substrate NucS^R. DMS reacts predominantly with A and C bases (RNA and FANA). Positions were defined as reactive (blue open bars and circles) if reactivity was greater than a cut-off (dashed line in **b** and **d**) of one half standard deviation above the median. Dashed circles indicate positions with marginal reactivity. Site of cleavage (in unmodified RNA) or ligation is indicated by a black inverted triangle. Primer-binding regions (no structural data) are shown in grey.



Extended Data Figure 6 | Selection scheme for RNA ligase XNAzymes.

a, XNA library preparation using DNA-dependent XNA polymerases, primed by a 5' triphosphorylated (5' ppp) RNA primer (LigS2^R), which serves as one of the substrates for RNA ligation *in cis*. Libraries are synthesized with 3' biotinylated DNA templates (N40temp(b); see Supplementary Table 1), allowing subsequent capture and removal by streptavidin beads. **b**, Single-stranded libraries (unbiotinylated) are annealed and incubated in reaction buffer (see Methods) together with a biotinylated chimaeric DNA–RNA substrate (tag1_LigS1^R), which successful XNAzymes ligate to RNA substrate LigS2^R *in cis*. **c**, Size separation of reacted XNA pools using urea–PAGE. Ligated

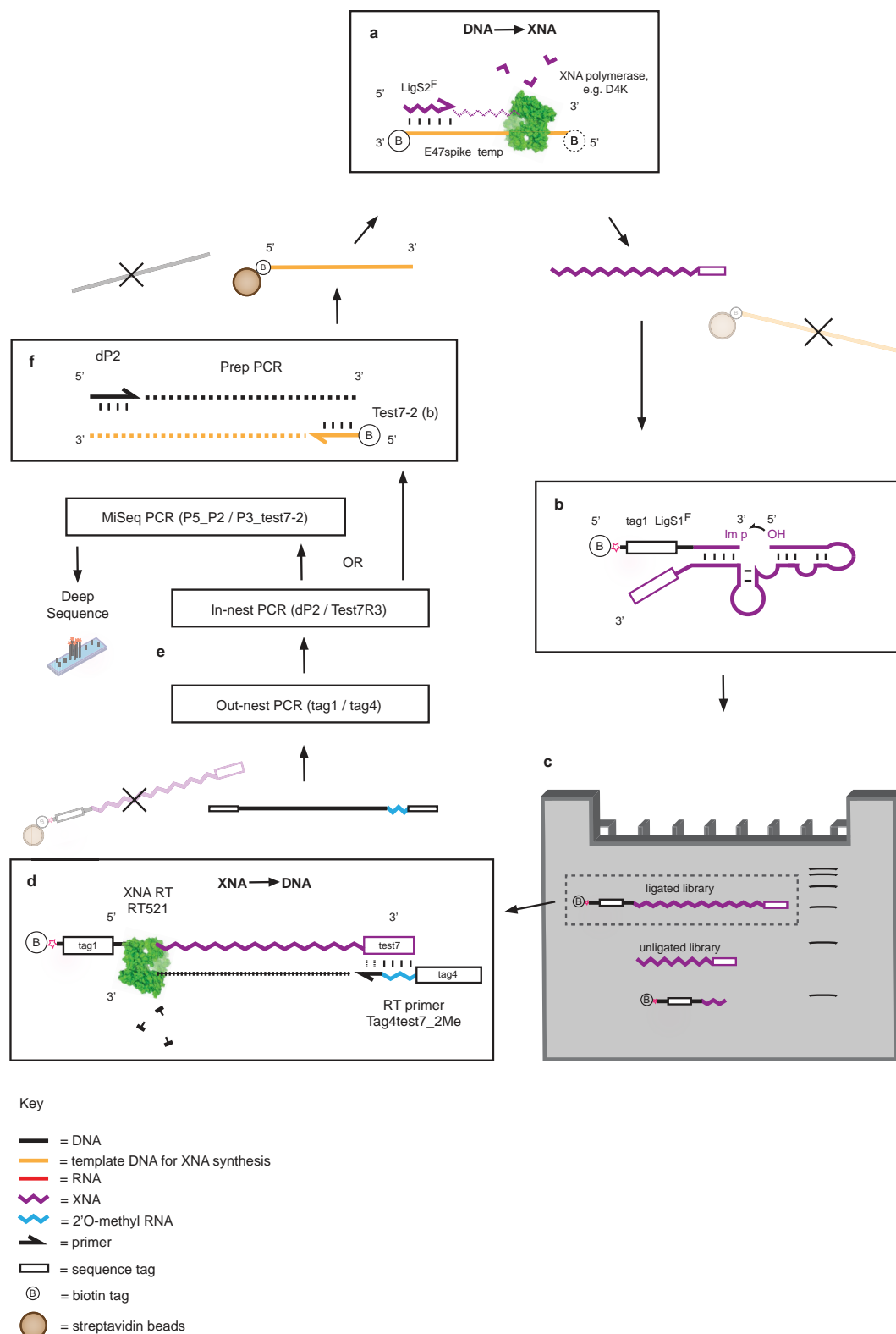
XNA pools are gel-extracted (indicated by dashed box). **d**, Reverse transcription of XNA pools using XNA-dependent DNA polymerase RT521L, which is also able to transcribe RNA across the ligation junction (that is, [RNA–RNA–XNA] \rightarrow cDNA). **e**, Amplification of transcribed cDNA by successive PCR reactions, using the primers indicated (see Supplementary Table 1); out-nest reaction depends on priming site (tag1) from ligated substrate tag1_LigS1^R. **f**, PCR reaction generating templates for XNA synthesis (now 5' biotinylated) for further rounds of selection. Solid crosses indicate removal of denatured strands using streptavidin bead capture.



Extended Data Figure 7 | Sequences and analyses of RNA ligase XNAzymes.

a, Schematic diagram showing RNA (red)–XNA (purple) chimaeric library setup for selection of FANAzymes capable of catalysing a bimolecular RNA ligation. **b**, Sequences of the FANA region under selection (dashed box in **a**) of the most abundant clones revealed by deep sequencing. Representatives of sequence families were screened for activity in bimolecular (LigS2^R attached to XNAzyme) or trimolecular (XNAzyme separate from both substrates) reactions. Sequence F2R17_1 (boxed) was chosen for further characterization. **c**, Regiospecificity of RNA product (LigP^R) of ligation catalysed by XNAzyme F2R17_1min (see Fig. 3), analysed by strong anion exchange chromatography

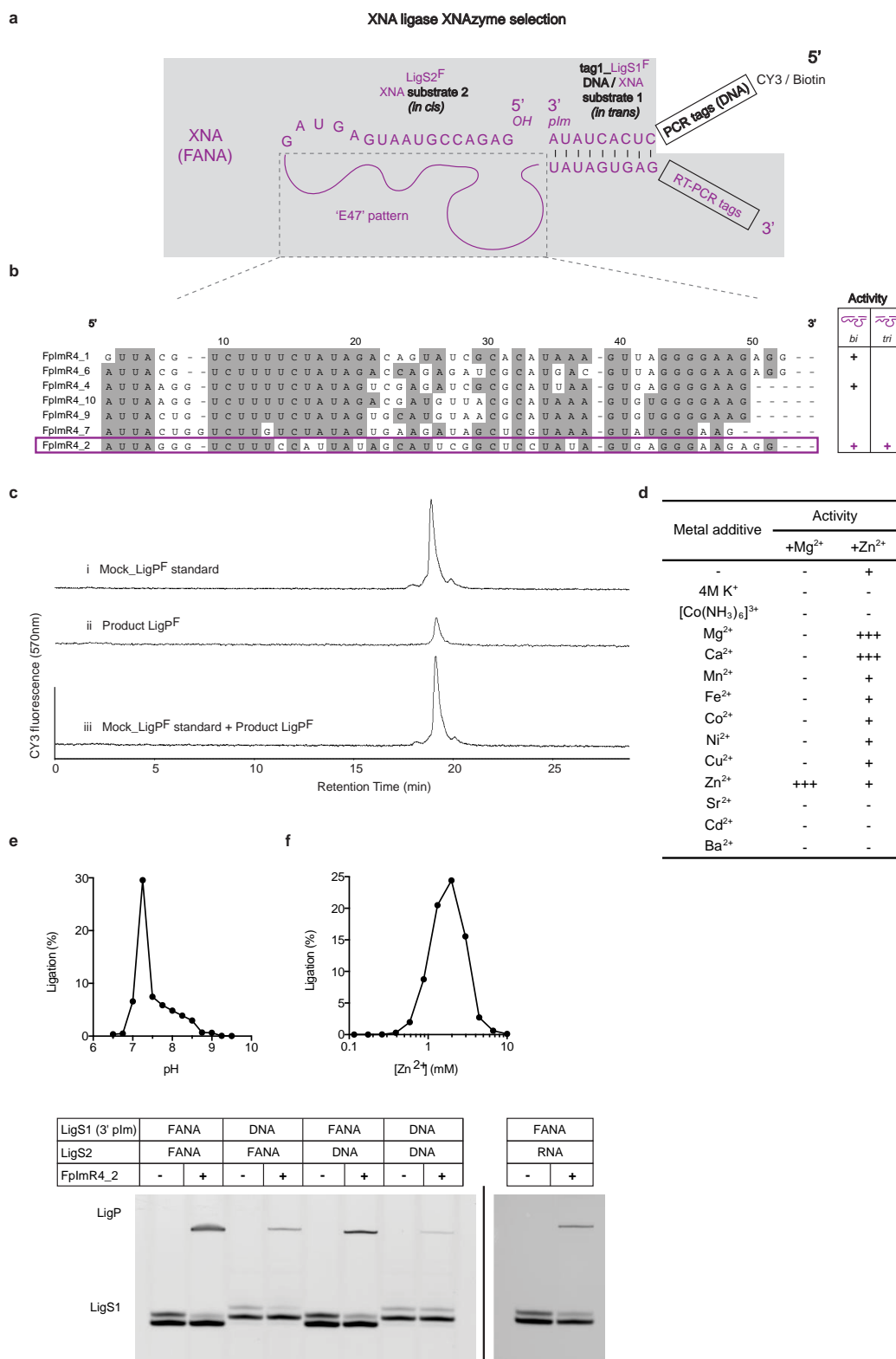
(SAX-HPLC)³⁶. Mock RNA ligation product (i–iii) containing a single 2′–5′ (Mock_LigP^R[2′–5′]) or 3′–5′ linkage (Mock_LigP^R[3′–5′]) at a position analogous to the ligation site were compared to the XNAzyme-catalysed RNA product LigP^R (iv–vi). The XNAzyme product gives an identical elution profile to the natural (3′–5′) linkage standard. **d**, Bivalent metal ion requirements and titration of, **e**, pH or **f**, MgCl₂, of FANAzyme F2R17_1min reaction. **g**, Substitution of RNA ligase substrates with DNA and XNA (FANA) versions in F2R17_1min reaction shows that 5′–RNA–RNA–3′ ligation is preferred, but some ligase activity can be seen with 5′–RNA–DNA–3′.



Extended Data Figure 8 | Selection scheme for XNA ligase XNAzymes.

a, XNA library preparation using DNA-dependent XNA polymerases, primed by an all-XNA (FANA) primer (LigS2^F), which serves as one of the substrates for FANA ligation in *cis*. Libraries are synthesized with 3' biotinylated DNA template (E47spike_temp; see Supplementary Table 1), allowing subsequent capture and removal by streptavidin beads. **b**, Single-stranded libraries (unbiotinylated) are annealed and incubated in reaction buffer (see Methods) together with a biotinylated chimaeric DNA–XNA (FANA) substrate (tag1_LigS1^F), activated with a 3' phosphorylimidazole (Extended Data Fig. 10), which successful XNAzymes ligate to XNA (FANA) substrate LigS2^F

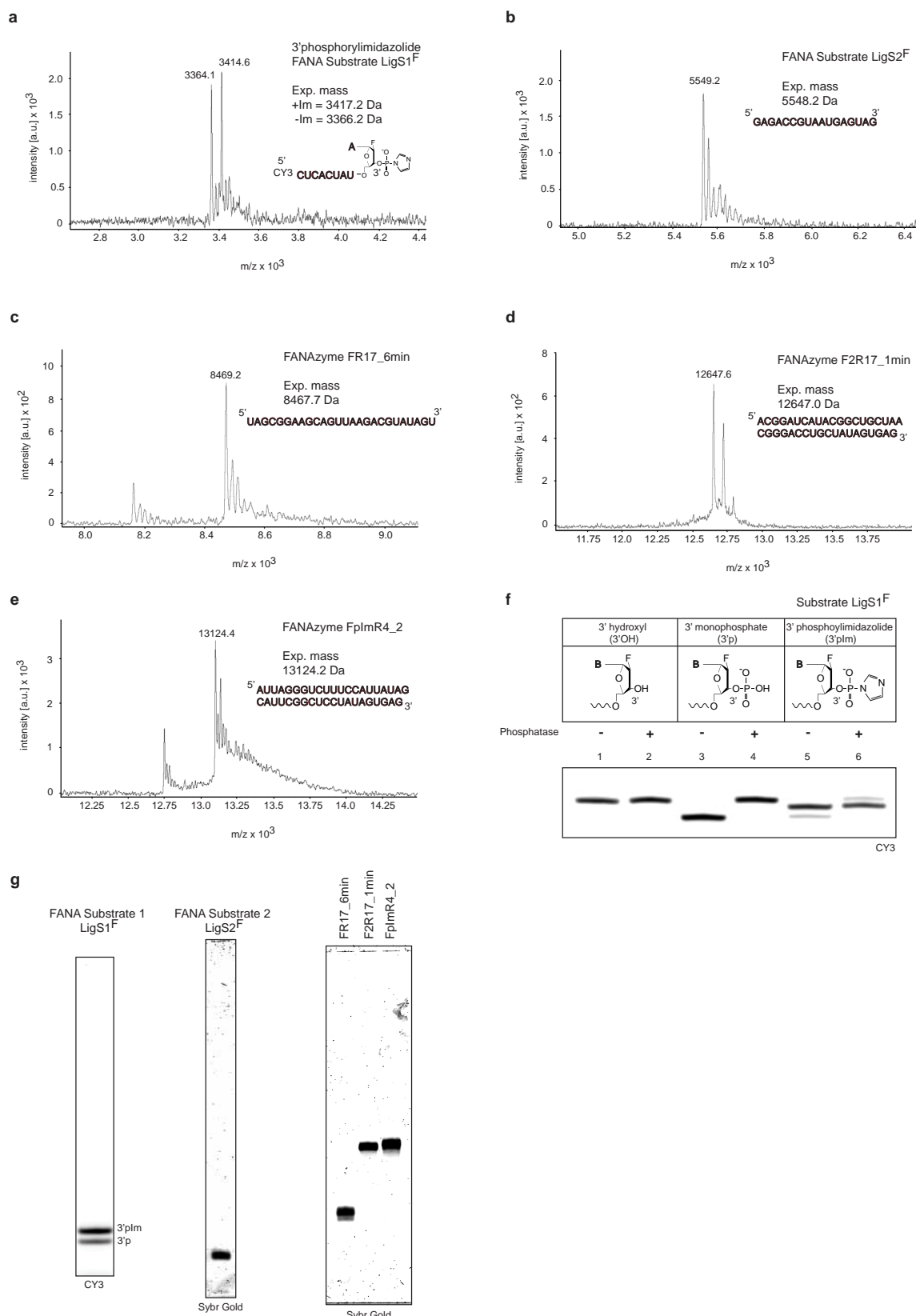
in *cis*. **c**, Size separation of reacted XNA pools using urea–PAGE. Ligated XNA pools are gel-extracted (indicated by dashed box). **d**, Reverse transcription of XNA pools using XNA-dependent DNA polymerase RT521L (that is, XNA → cDNA). **e**, Amplification of transcribed cDNA by successive PCR reactions, using the primers indicated (see Supplementary Table 1); out-nest reaction depends on priming site (tag1) from ligated substrate tag1_LigS1^F. **f**, PCR reaction generating templates for XNA synthesis (now 5' biotinylated) for further rounds of selection. Solid crosses indicate removal of denatured strands using streptavidin bead capture.



Extended Data Figure 9 | Sequences and analyses of XNA ligase XNAzymes (FANA). **a**, Schematic diagram showing all-FANA library setup for selection of FANAzymes capable of catalysing a bimolecular XNA (FANA) ligation.

b, FANA sequences of the region under selection (dashed box in **a**) of the most abundant clones revealed by deep sequencing. Representatives of sequence families were screened for activity in bimolecular (LigS2^F attached to XNAzyme) or trimolecular (XNAzyme separate from both substrates) reactions. Sequence FplmR4_2 (boxed) was chosen for further characterization. **c**, Regiospecificity of XNA (FANA) product (LigP^F) of ligation catalysed by XNAzyme FplmR4_2 (see Fig. 4), analysed by strong

anion exchange chromatography (SAX-HPLC). Mock FANA ligation product (Mock_LigP^F) (i), prepared by polymerase (D4K) gives an identical elution profile to the XNAzyme-catalysed FANA product (LigP^F) (ii and iii). **d**, Bivalent metal ion requirements and titration of, **e**, pH or **f**, MgCl₂, of FANAzyme FplmR4_2 reaction. **g**, Substitution of XNA ligase substrates with RNA and DNA versions in FplmR4_2 reaction. Although 5'-FANA-FANA-3' is preferred, ligase activity can be seen with 5'-FANA-DNA-3', and 5'-FANA-RNA-3', as well as 5'-DNA-FANA-3', 5'-DNA-DNA-3' and 5'-DNA-RNA-3'.



Extended Data Figure 10 | Analysis of XNA (FANA) substrates and enzymes prepared by solid-phase synthesis. MALDI-ToF mass spectra showing expected masses of **a**, XNA (FANA) ligase substrate LigS1^F-3'phosphorylimidazolide (prepared by solid-phase synthesis of the 3' phosphorylated (3'p) oligonucleotide, followed by reaction with carbodiimide and imidazole (see Methods)), **b**, XNA (FANA) ligase substrate LigS2^F, and XNAzymes **c**, FR17_6min, **d**, F2R17_6min, and **e**, FpImR4_2.

f, Dephosphorylation assay of versions of LigS1^F (3' hydroxyl, lanes 1 and 2, 3' phosphate, lanes 2 and 3, or 3' phosphorylimidazolide, lanes 5 and 6) with calf intestinal phosphatase (lanes 2, 4 and 6). The majority of the LigS1^F preparation shown (~70%) is protected from dephosphorylation, consistent with formation of the 3' plm. **g**, Urea-PAGE analyses of purified FANA substrates and XNAzymes.

Structure of the key species in the enzymatic oxidation of methane to methanol

Rahul Banerjee^{1,2}, Yegor Proshlyakov³, John D. Lipscomb^{1,2} & Denis A. Proshlyakov³

Methane monooxygenase (MMO) catalyses the O₂-dependent conversion of methane to methanol in methanotrophic bacteria, thereby preventing the atmospheric egress of approximately one billion tons of this potent greenhouse gas annually. The key reaction cycle intermediate of the soluble form of MMO (sMMO) is termed compound Q (Q). Q contains a unique dinuclear Fe^{IV} cluster that reacts with methane to break an exceptionally strong 105 kcal mol⁻¹ C-H bond and insert one oxygen atom^{1,2}. No other biological oxidant, except that found in the particulate form of MMO, is capable of such catalysis. The structure of Q remains controversial despite numerous spectroscopic, computational and synthetic model studies²⁻⁷. A definitive structural assignment can be made from resonance Raman vibrational spectroscopy but, despite efforts over the past two decades, no vibrational spectrum of Q has yet been obtained. Here we report the core structures of Q and the following product complex, compound T, using time-resolved resonance Raman spectroscopy (TR³). TR³ permits fingerprinting of intermediates by their unique vibrational signatures through extended signal averaging for short-lived species. We report unambiguous evidence that Q possesses a bis-μ-oxo diamond core structure and show that both bridging oxygens originate from O₂. This observation strongly supports a homolytic mechanism for O-O bond cleavage. We also show that T retains a single oxygen atom from O₂ as a bridging ligand, while the other oxygen atom is incorporated into the product⁸. Capture of the extreme oxidizing potential of Q is of great contemporary interest for bioremediation and the development of synthetic approaches to methane-based alternative fuels and chemical industry feedstocks. Insight into the formation and reactivity of Q from the structure reported here is an important step towards harnessing this potential.

Transient kinetic studies of sMMO have revealed eight reaction cycle intermediates, thereby providing the most comprehensive description of enzymatic O₂ activation and C-H bond oxidation currently available for any di-iron oxygenase^{1,9-11} (Fig. 1a). It is broadly accepted that the linear increase in the decay rate constant for Q with concentration of methane indicates reaction between Q and substrates with concomitant formation of the product complex T. Q is formed in a single turnover system by mixing O₂-containing buffer solution with the heterocomplex of diferrous sMMO hydroxylase (MMOH^{red}) and regulatory B component (MMOB). In the absence of methane, the yellow Q has a lifetime of several seconds, accumulates in high yield, and can be trapped by rapid freeze-quench (RFQ) techniques. Several spectroscopic studies of trapped Q have been successful^{2,3,10}, but not RFQ-resonance Raman, perhaps owing to weak resonance Raman enhancement and/or photosensitivity of Q. To minimize photolysis, we acquired the resonance Raman spectrum of Q in a continuously flowing reactant stream, while at the same time extending spectral accumulation to many hours¹². Previous studies have shown that Q maximizes at Δt ≈ 3 s after initiation of the reaction between MMOH^{red}/MMOB and O₂ at pH 7.0, 4 °C in the absence of substrate¹ (Extended Data Fig. 1a). Accordingly, we observe the electronic absorption spectrum of Q at this time in the TR³ flow cell (Fig. 1b).

The absolute resonance Raman spectra of the reaction mixture at Δt ≈ 3 s (λ_{ex} = 351 nm) are dominated by non-resonant vibrations of bulk solution due to the relatively low attainable enzyme concentration (0.25 mM) (Extended Data Fig. 2). Repetitive switching between ¹⁶O₂- and ¹⁸O₂-saturated oxygen streams, while maintaining the same differential MMOH^{red}/MMOB stream, eliminates minute variability between sample preparations. The O₂ isotope difference spectra reveal weak vibrations that involve dioxygen-derived atoms, thus identifying sMMO intermediates, while all other vibrations cancel out (Fig. 1c, trace (i), and Extended Data Fig. 2).

Two distinct oxygen vibrations were detected at Δt ≈ 3 s in the absence of substrate (Fig. 1c, trace (i)): a major mode at 690 cm⁻¹ (identified herein by the ¹⁶O isotopomer and the ¹⁶O/¹⁸O downshift, Δ¹⁸O = 36 cm⁻¹) and a weaker mode at 556 cm⁻¹ (Δ¹⁸O = 23 cm⁻¹). No

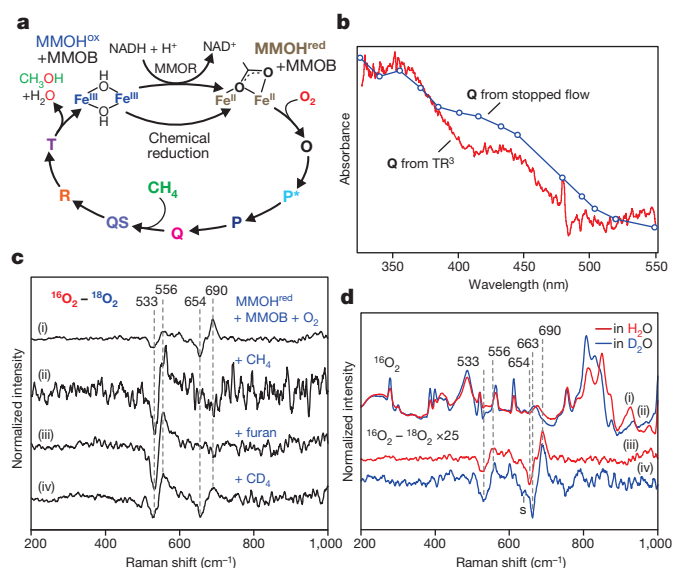


Figure 1 | Reaction of sMMO with O₂. **a**, The catalytic cycle includes stable MMOH^{ox} and MMOH^{red} and detectable transient species O, P*, P, Q and T; transient states QS and R were predicted from kinetic, spectroscopic, and chemical studies. **b**, An *in situ* transient difference electronic absorption spectrum of the reaction mixture (vs anaerobic MMOH^{red}) in the TR³ instrument at Δt ≈ 3.0 s (red, 1.0 × 0.1² mm probe volume) in comparison with spectrum of Q reconstructed from stopped-flow kinetic traces (markers, blue). **c**, Transient ¹⁶O₂ - ¹⁸O₂ difference resonance Raman spectra of sMMO reveal vibrations of iron-bound oxygen atoms. Measurement conditions: pH 7.0, 4 °C, Δt ≈ 3.0 s. Spectra recorded without the substrate (i) show marked changes when 0.45 mM CH₄ (ii), 3.5 mM furan (iii) or 0.45 mM CD₄ (iv) is added. **d**, Solvent vibrations in the absolute resonance Raman spectra are sensitive to H₂O (i)/D₂O (ii) substitution. ¹⁶O₂ - ¹⁸O₂ difference spectra of sMMO recorded in H₂O (iii) show little sensitivity to D₂O (iv) substitution. The upshift of ¹⁸O vibration of Q and the appearance of a low-frequency shoulder (marked with S) in D₂O is attributed to Fermi resonance with a protein-derived metal ligand.

¹Department of Biochemistry, Molecular Biology & Biophysics, University of Minnesota, Minneapolis, Minnesota 55455, USA. ²Center for Metals in Biocatalysis, University of Minnesota, Minneapolis, Minnesota 55455, USA. ³Department of Chemistry, Michigan State University, East Lansing, Michigan 48824, USA.

Table 1 | Vibrational modes of metal centres relevant to Q and T

Vibrational structure†	System	$\nu(\Delta^{18}\text{O}) \text{ cm}^{-1}$	$\nu^{16}\text{O}^{18}\text{O} \text{ cm}^{-1}$	$\Delta^2\text{H} \text{ cm}^{-1}$	Ref.
	[4-OMe-3,5-Me-TPAFe ^{IV} (μ-O)] ₂	674 (–30)	665	–	7
	[(TPA) ₂ Fe ^{III} (μ-O) ₂ Fe ^{IV}]	666 (–28)*	644	–	19
	[(5-Me-TPA) ₂ Fe ³⁺ (μ-O) ₂ Fe ^{IV}]	668 (–32)	–	–	19
	sMMO Q	690 (–36)	673	0	This work
	sMMO T	556 (–23)	–	0	This work
	[Fe ₂ (μ-O)(μ-OH)(4,6-Me ₆ -TPA) ₂] ^{III}	497 (–13)	–	–3	25
	[Fe ₂ (μ-O)(μ-OH)(BPEEN) ₂] ^{III}	504 (–16)	–	–4	25
	Ribonucleotide reductase-R2	496 (–15)	–	0	26
	Δ ⁹ -desaturase	519 (–18)	–	0	27
	Urease UreA2B2	497 (–21)	–	–	28
	Hydroxomet-hemerythrin	565 (–27)	–	–5	16
	Chloroperoxidase compound-II	565 (–22)	–	–13	17
	Horseradish peroxidase	503 (–19)	–	+6	29

*Represents centre of a Fermi doublet.

†Additional cases are considered in references 12, 15 and as noted in Extended Data Table 1.

noticeable laser power dependence was observed, thus excluding photochemistry under these conditions (Extended Data Fig. 3). Neither of the vibrations was observed at a long delay time ($\Delta t \approx 30$ s).

The catalytic relevance of the observed intermediate(s) was probed by mixing substrates into the reactant stream, which is expected to fully quench Q¹. Accordingly, the 690 cm^{–1} mode disappeared upon addition of methane or furan (Fig. 1c, traces (ii) and (iii)). A concomitant fourfold increase in the intensity of the 556 cm^{–1} mode shows that this vibration arises from a different intermediate that evolves from Q as it reacts with substrate. The rate constants for Q decay with methane or furan predict a predominant accumulation of T at $\Delta t \approx 3$ s (ref. 1; Extended Data Fig. 1b, c). A broad near-ultraviolet electronic absorption band for T allows its resonance enhancement ($\lambda_{\text{ex}} = 351$ nm) (Extended Data Fig. 4). Based on substrate-induced build-up and its transient

nature, the 556 cm^{–1} vibration mode is assigned to T. The observation of the resonance Raman spectrum of T in the absence of substrate suggests that the spontaneous decay of Q involves similar intermediates as found in the substrate mediated process.

The exceptionally large deuterium kinetic isotope effect in the reaction of Q with methane⁸ reveals further correlation between Q and the 690 cm^{–1} mode. This step is slowed by a factor of 50 when using CD₄ rather than CH₄ due to the large tunnelling component in the reaction coordinate¹³ (Extended Data Fig. 1d). Indeed, the loss of intensity of the 690 cm^{–1} vibration and the increase of the 556 cm^{–1} vibration were much smaller when using CD₄ as substrate (Fig. 1c, trace (iv)).

An initial assignment of the structure for the di-iron cluster of Q can be made by comparing its resonance Raman spectrum to those of model complexes (Table 1, Extended Data Table 1). The observed frequency

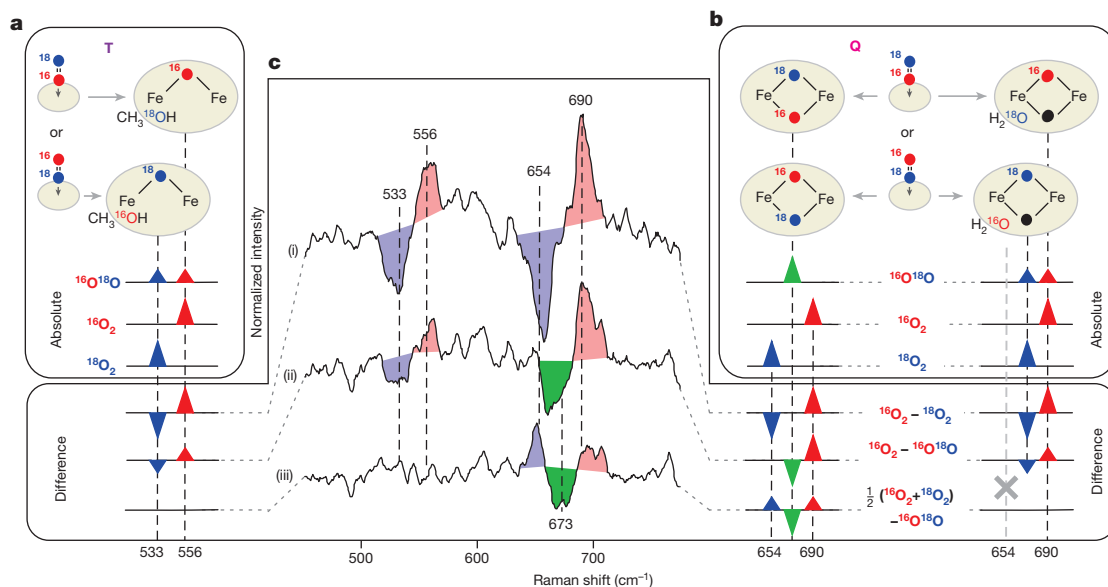


Figure 2 | Fingerprinting cluster structure using ¹⁶O¹⁸O mixed oxygen isotope. **a, b,** Asymmetrically labelled ¹⁶O¹⁸O can initially bind in two equiprobable orientations, yielding an even mixture of two isotopomers in T or Q (panels **a** and **b**, top) exhibiting characteristic vibrations (panels **a** and **b**, bottom). Two scenarios are possible for Q, depending on whether two (left) or one (right) O₂-derived atoms (red and blue) are incorporated into the cluster. If only one atom is incorporated, then the second oxygen atom in the core structure would derive from solvent (black). Water and methanol exemplify a departing oxygen atom while other ligands are omitted for simplicity. Since the two isotopomers of T (**a**) and those of singly labelled Q (**b**, right) have identical composition as corresponding ¹⁶O₂ and ¹⁸O₂ derivatives, they will exhibit both vibrations simultaneously at half the

intensity. Doubly labelled Q (panel **b**, left) will be different from both symmetrically labelled derivatives and thus, **c**, will exhibit a new vibration (green), as illustrated by isotope difference spectra (**c**). The ¹⁶O₂ – ¹⁶O¹⁸O difference (ii) in singly labelled cluster (**a**, **b** right) will appear as the ¹⁶O₂ – ¹⁸O₂ difference (i) with reduced intensity. The ¹⁶O¹⁸O derivative will be identical to the average of symmetrical isotopomers, yielding no signal in trace iii, as observed experimentally for T. The new frequency in doubly labelled Q should appear in both difference (ii) and (iii), and can, indeed, be seen in experimental data at 673 cm^{–1}. Spectral superimposition inflates the apparent isotopic shift when frequencies of isotopomers are close (traces (ii) and (iii), right), but not when bands are well separated (trace (i)).

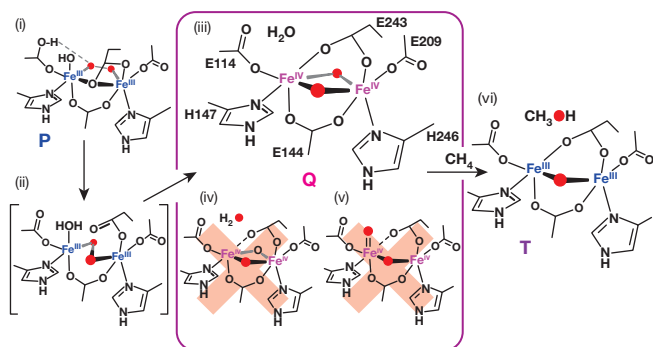


Figure 3 | Formation of compound Q and its reaction with methane.

Homolytic mechanism of O–O bond cleavage (ii) upon formation of Q from P (i) follows from the structure of Q (iii) presented here. Alternative structures of Q (iv, v) discounted by current results are also shown. The iron shown on the left in the Q structure (iii) may retain the solvent found in P. In this case, E243 would not bind to this iron, but would be likely to hydrogen bond with the bound solvent^{4,5}. T (vi) contains a single atom from O₂ while another is incorporated in the product.

of 690 cm^{−1} for Q is much less than the $\nu_{\text{Fe}=\text{O}}$ of terminal ferryl complexes^{12,14}. Moreover, the $\Delta^{18}\text{O}$ shift of 36 cm^{−1} for Q is greater than the expected value for an Fe^{IV}=O diatomic oscillator at this frequency. Therefore, the possibility of an isolated, terminal Fe^{IV}=O moiety in Q can be excluded (Fig. 3, panel (v)), as can be a peroxy complex, which would exhibit the $\nu_{\text{O}-\text{O}}$ at even higher frequency¹⁵. In contrast, the $\nu_{\text{Fe}-\text{O}}$ of end-on Fe^{III}-OH, Fe^{IV}-OH and Fe^{II}-O₂ complexes appear at lower frequencies than observed here^{16–18}. Protonated ligands typically exhibit a pronounced downshift upon bulk water deuteration, which we did not observe for Q (Fig. 1d). Indeed, the vibration frequency and $\Delta^{18}\text{O}$ isotopic shift for Q agree with values observed in only one class of synthetic di-iron models, namely bis- μ -oxo ‘diamond core’ complexes¹⁹.

The defining feature of a diamond core structure is the presence of two vibrationally coupled μ -oxo bridges resulting in tetraatomic vibrations (Table 1). This coupled motion yields a unique signature that allows discrimination of the diamond core from other structures by using the mixed isotopomer of oxygen (¹⁶O¹⁸O) (ref. 19). If the diamond core comprises both oxygen atoms from O₂, then mixed oxygen substitution in Q (Q-¹⁶O¹⁸O) will result in a new vibration appearing symmetrically between those of the Q-¹⁶O₂ and Q-¹⁸O₂ (Fig. 2b, c, green). However, the absence of a diamond core or the incorporation of only one of two oxygen atoms from O₂ (Fig. 3, panel (iv)) will yield the same vibrational frequencies as observed for Q-¹⁶O₂ and Q-¹⁸O₂ with smaller $\Delta^{18}\text{O}$ (Fig. 2a–c, blue, red). Subtraction of $\nu(\text{Q-}^{16}\text{O}^{18}\text{O})$ from the average of $\nu(\text{Q-}^{16}\text{O}_2)$ and $\nu(\text{Q-}^{18}\text{O}_2)$ (Fig. 2c, trace (iii)) reveals a new frequency at ~673 cm^{−1} in the resonance Raman spectrum of the asymmetrically labelled derivative, showing definitively that it arises from a diamond core di-iron cluster of Q. Notably, the new vibration also indicates that Q retains both oxygen atoms from O₂. The absence of Q-¹⁶O¹⁸O vibration in the ¹⁶O₂–¹⁸O₂ difference (Fig. 2c, trace (i)) shows that no exchange occurs between O₂ atoms and bulk water (Extended Data Fig. 5).

Subtraction of $\nu(\text{T-}^{16}\text{O}^{18}\text{O})$ from the average of $\nu(\text{T-}^{16}\text{O}_2)$ and $\nu(\text{T-}^{18}\text{O}_2)$ results in cancellation of all oxygen vibrations (Fig. 2c, trace (iii)), demonstrating that T inherits only one of two oxygen atoms from O₂. This vibration is unchanged for samples prepared in D₂O (Fig. 1d), strongly suggesting that T does not possess a protonated bridging oxygen or terminal Fe–OH moiety, which are the only other species with similar resonance Raman vibrational fingerprints (Table 1). Thus, we assign T as a mono- μ -oxo-bridged structure derived from Q by transfer of one oxygen atom to product, as observed experimentally⁸ (Fig. 3, panel (vi)). We found no indication that labelled product is directly bound to the di-iron cluster in T. Together with the Q data, these observations allow each oxygen atom from O₂ to be tracked throughout the reaction cycle.

The observation that both bridging atoms in Q are derived from the same O₂ molecule has important implications for the mechanism of O₂ activation in sMMO. The dinuclear iron cluster in the P intermediate (Fig. 3, panel (i)) that precedes Q is predicted to possess a symmetrical *cis*- μ -1,2 peroxo bridge²⁰. Proton migration in P facilitates O–O bond cleavage to form Q, which could occur either heterolytically (as in cytochrome P450)²¹ or homolytically (as in some dinuclear copper–O₂ complexes and mimics of NO reductase)^{22,23}. A heterolytic cleavage mechanism would result in formation of a water molecule from an O₂-derived atom (Fig. 3, panel (iv), and Extended Data Fig. 5c). Our results show that neither of the O₂-derived atoms is lost from the cluster, nor do we see vibrations of O₂-derived water bound to the iron atoms. While it is possible to formulate a mechanism for symmetrical diamond core formation by heterolytic O–O cleavage (Extended Data Fig. 5b), such a scenario would involve large asymmetric changes in the oxidation states of the cluster irons and would probably show solvent oxygen incorporation into the diamond core of Q, which we do not observe. Consequently, O₂ activation in sMMO is likely to proceed via homolytic cleavage (Fig. 3, panels (ii) and (iii), and Extended Data Fig. 5a), highlighting different mechanistic strategies of haem and non-haem enzymes for generation of potent oxygenating species.

A previous combined Mössbauer and extended X-ray absorption fine structure study of Q revealed an exceptionally short Fe–Fe distance³. The model complexes available at the time suggested that the short Fe–Fe distance could be accommodated by a bis- μ -oxo ‘diamond-core’ structure³. In recent years, other structures of Q have been proposed based on new synthetic model complexes^{20,24}. It was found that formation of an Fe^{IV}=O moiety in an open-core low-spin Fe^{IV}/Fe^{IV} or high-spin Fe^{III}/Fe^{IV} complex resulted in major increases in reactivity compared with the low-spin closed diamond core Fe^{IV}/Fe^{IV} or Fe^{III}/Fe^{IV} models with the same ligand structure²⁴. Contrary to several open-core proposals for Q²⁰ that developed from this finding (Fig. 3, panel (v)), we do not detect an Fe^{IV}=O moiety indicative of an open core in Q. A coupled, polyatomic vibration in Q shows that it does have a diamond-core structure. Consequentially, the key to the reactivity of nature’s most potent oxidant is likely to come from the spin state of the cluster, putting emphasis on further synthetic work towards mimicking the high-spin nature of Q, while retaining the high-valent diamond core.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 14 September; accepted 22 December 2014.

Published online 21 January 2015.

- Lee, S. K., Nesheim, J. C. & Lipscomb, J. D. Transient intermediates of the methane monooxygenase catalytic cycle. *J. Biol. Chem.* **268**, 21569–21577 (1993).
- Lee, S. K., Fox, B. G., Froland, W. A., Lipscomb, J. D. & Münck, E. A transient intermediate of the methane monooxygenase catalytic cycle containing a Fe^{IV}/Fe^{IV} cluster. *J. Am. Chem. Soc.* **115**, 6450–6451 (1993).
- Shu, L. *et al.* An Fe^{IV}₂O₂ diamond core structure for the key intermediate Q of methane monooxygenase. *Science* **275**, 515–518 (1997).
- Dunietz, B. D. *et al.* Large scale ab initio quantum chemical calculation of the intermediates in the soluble methane monooxygenase catalytic cycle. *J. Am. Chem. Soc.* **122**, 2828–2839 (2000).
- Han, W. G. & Noodleman, L. Structural model studies for the high-valent intermediate Q of methane monooxygenase from broken-symmetry density functional calculations. *Inorganica Chim. Acta* **361**, 973–986 (2008).
- Siegbahn, P. E. M. O–O bond cleavage and alkane hydroxylation in methane monooxygenase. *J. Biol. Inorg. Chem.* **6**, 27–45 (2001).
- Xue, G. *et al.* A synthetic precedent for the [Fe^{IV}₂(μ -O)₂] diamond core proposed for methane monooxygenase intermediate Q. *Proc. Natl Acad. Sci. USA* **104**, 20713–20718 (2007).
- Nesheim, J. C. & Lipscomb, J. D. Large isotope effects in methane oxidation catalyzed by methane monooxygenase: evidence for C–H bond cleavage in a reaction cycle intermediate. *Biochemistry* **35**, 10240–10247 (1996).
- Liu, K. E. *et al.* Spectroscopic detection of intermediates in the reaction of dioxygen with the reduced methane monooxygenase hydroxylase from *Methylococcus capsulatus* (Bath). *J. Am. Chem. Soc.* **116**, 7465–7466 (1994).
- Tinberg, C. E. & Lippard, S. J. Revisiting the mechanism of dioxygen activation in soluble methane monooxygenase from *M. capsulatus* (Bath): evidence for a multi-step, proton-dependent reaction pathway. *Biochemistry* **48**, 12145–12158 (2009).

11. Banerjee, R., Meier, K. K., Münck, E. & Lipscomb, J. D. Intermediate P* from soluble methane monooxygenase contains a diferrous cluster. *Biochemistry* **52**, 4331–4342 (2013).
12. Grzyska, P. K., Appelman, E. H., Hausinger, R. P. & Proshlyakov, D. A. Insight into the mechanism of an iron dioxygenase by resolution of steps following the Fe^{IV}=O species. *Proc. Natl Acad. Sci. USA* **107**, 3982–3987 (2010).
13. Zheng, H. & Lipscomb, J. D. Regulation of methane monooxygenase catalysis based on size exclusion and quantum tunneling. *Biochemistry* **45**, 1685–1692 (2006).
14. Hohenberger, J., Kallol, R. & Meyer, K. K. The biology and chemistry of high-valent iron-oxo and iron-nitrido complexes. *Nat. Commun.* **3**, 720 (2012).
15. Vu, V. V. *et al.* Human deoxyhypusine hydroxylase, an enzyme involved in regulating cell growth, activates O₂ with a nonheme diiron center. *Proc. Natl Acad. Sci. USA* **106**, 14814–14819 (2009).
16. Shiemke, A. K., Loehr, T. M. & Sanders-Loehr, J. Resonance Raman study of oxyhemerythrin and hydroxhemerythrin: evidence for hydrogen bonding of ligands to the Fe–O–Fe center. *J. Am. Chem. Soc.* **108**, 2437–2443 (1986).
17. Stone, K. L., Behan, R. K. & Green, M. T. Resonance Raman spectroscopy of chloroperoxidase compound II provides direct evidence for the existence of an iron(IV)-hydroxide. *Proc. Natl Acad. Sci. USA* **103**, 12307–12310 (2006).
18. Momenteau, M. & Reed, C. A. Synthetic heme dioxygen complexes. *Chem. Rev.* **94**, 659–698 (1994).
19. Wilkinson, E. C. *et al.* Raman signature of the Fe₂O₂ “diamond” core. *J. Am. Chem. Soc.* **120**, 955–962 (1998).
20. Tinberg, C. E. & Lippard, S. J. Dioxygen activation in soluble methane monooxygenase. *Acc. Chem. Res.* **44**, 280–288 (2011).
21. Poulos, T. L. Heme enzyme structure and function. *Chem. Rev.* **114**, 3919–3962 (2014).
22. Tolman, W. B. Making and breaking the dioxygen O–O bond: new insights from studies of synthetic copper complexes. *Acc. Chem. Res.* **30**, 227–237 (1997).
23. Collman, J. P., Dey, A., Yang, Y., Ghosh, S. & Decreau, R. A. O₂ reduction by a functional heme/nonheme bis-iron NOR model complex. *Proc. Natl Acad. Sci. USA* **106**, 10528–10533 (2009).
24. Xue, G., De Hont, R., Münck, E. & Que, L. Jr. Million-fold activation of the [Fe₂(μ-O)₂] diamond core for C–H bond cleavage. *Nat. Chem.* **2**, 400–405 (2010).
25. Zheng, H., Zang, Y., Dong, Y., Young, V. G. Jr & Que, L. Jr. Complexes with Fe^{III}₂(μ-O)(μ-OH), Fe^{III}₂(μ-O)₂, and [Fe^{III}₃(μ₂-O)₃] cores: Structures, spectroscopy, and core interconversions. *J. Am. Chem. Soc.* **121**, 2226–2235 (1999).
26. Sjöberg, B.-M., Loehr, T. M. & Sanders-Loehr, J. Raman spectral evidence for a μ-oxo bridge in the binuclear iron center of ribonucleotide reductase. *Biochemistry* **21**, 96–102 (1982).
27. Fox, B. G., Shanklin, J., Ai, J., Loehr, T. M. & Sanders-Loehr, J. Resonance Raman evidence for an Fe–O–Fe center in stearyl-ACP desaturase. Primary sequence identity with other diiron-oxo proteins. *Biochemistry* **33**, 12776–12786 (1994).
28. Carter, E. L., Proshlyakov, D. A. & Hausinger, R. P. Apoprotein isolation and activation, and vibrational structure of the *Helicobacter mustelae* iron urease. *J. Inorg. Biochem.* **111**, 195–202 (2012).
29. Sitter, A. J., Shifflott, J. R. & Terner, J. Resonance Raman spectroscopic evidence for heme iron-hydroxide ligation in peroxidase alkaline forms. *J. Biol. Chem.* **263**, 13032–13038 (1988).

Acknowledgements We thank G. T. Babcock (deceased), S.-K. Lee and J. C. Nesheim (deceased) for initial studies that led to this project and E. Bergeron for technical assistance. This work was supported by the NIH grants GM40466 and GM100943 (to J.D.L.) and grant GM096132 (to D.A.P.).

Author Contributions R.B., Y.P. and D.A.P. developed the enhanced instrument and performed the experiments, D.A.P. analysed the data, and R.B., D.A.P. and J.D.L. designed the experiments and wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.A.P. (dapro@chemistry.msu.edu) or J.D.L. (lipsc001@umn.edu).

METHODS

No statistical methods were used to predetermine sample size.

Chemicals. All chemicals used were of the highest grade available and were purchased from Sigma-Aldrich. $^{18}\text{O}_2$ was obtained from ICON. The asymmetrically labelled oxygen isotope ($^{16}\text{O}^{18}\text{O}$) was synthesized by E. H. Appelman from Argonne National Laboratory¹². D_2O was obtained from Cambridge Isotope Laboratories.

Biological materials. MMOH was purified from *Methylosinus trichosporium* OB3b and MMOB from a heterologous expression system in *Escherichia coli* according to purification protocols described previously^{11,30}. Low molar extinction of transient intermediates compounded by the nature of the TR^3 experiment required a large amount of the MMOH protein, estimated to be about 30 g for the experimental results presented here, with more than half used in the instrumentation development and optimization phases.

Sample preparation. Anaerobic solutions of MMOH in 50 mM MOPS, pH 7.0, 5% glycerol and MMOB in 50 mM MOPS, pH 7.0 were prepared in sealed vials by exchanging the headspace gas with argon while stirring on ice for 1 h. The chemical reduction of MMOH was conducted by anaerobic addition of methyl viologen (200 μM) and sodium hydrosulphite (4 mM) while stirring for 15 min at room temperature. MMOH was subsequently separated from the small molecule reducing agents by passage through a Sephadex G-25 PD-10 desalting column equilibrated in 75 mM MOPS buffer at pH 7.0 containing 5% glycerol, 0.1% Triton X-100, 200 μM $\text{Fe}(\text{NH}_4)_2\text{SO}_4$ and 2.0 mM L-cysteine (buffer A). The glycerol stabilizes the diferrous enzyme while Triton X-100 prevents protein aggregation in the flow cuvette that leads to disruption of the laser beam. The exogenous iron included in Buffer A in the form of the Fe^{II} -cysteine complex is used to maintain MMOH activity and does not reconstitute the protein active site³¹. Extensive controls have been carried out to ensure that it does not contribute to the observed vibrational spectra. Control stopped-flow experiments showed that neither glycerol nor Triton X-100 affected the decay rate constant for Q. Anaerobic MMOB was added to MMOH^{red} to obtain a 1:1 MMOH active sites to MMOB ratio (480 μM). The protein was transferred to the secondary glove bag (Plas-Labs) where it was installed on the syringe infusion pump. The running enzyme syringe was continuously chilled to 4 °C using a Peltier cooling block mounted atop the syringe since MMOH^{red} slowly precipitates at room temperature during the time required for a measurement (typically 1–5 h). For the experiments with deuterated solvents, the MMOH sample was exchanged by passage through the PD-10 desalting column equilibrated in buffer A prepared in D_2O (pH reading was adjusted to correct for D^+ vs H^+ concentration). The isotopic enrichment of deuterated samples is estimated at >90%. The TR^3 experiment was conducted at 5–6 °C to avoid freezing of the D_2O buffer.

TR^3 measurements. While the core design of the instrument is similar to the cryogenic TR^3 study described previously¹², a brief overview of the instrument and experimental design are provided to describe the extensive revisions that were required for the current study. The TR^3 experiment relies on active mixing of the continuous streams of MMOH^{red} /MMOB and oxygen-saturated buffer at a volumetric ratio of 1:1. The maximal concentration of Q in the absence of substrate was estimated to be 95 μM at the probe point, which arises from the 40% active fraction of MMOH^{II} . Sample is delivered to a fused silica flow cuvette (25 mm long, internal cross section 0.15×1.5 mm, Starna Cells) where it is probed approximately 5 mm from the mixer outlet (dead volume ~ 3 μl). The time resolution is achieved by varying the flow rates of the mixed stream driven by computer controlled syringe infusion pumps (high-pressure OEM modules, Harvard Apparatus). Flow rates varying between 100 μl per min and 2.5 μl per min per line result in ageing times of 1–36 s. The mixer with the attached flow cuvette are mounted on a translation stage driven by an actuator drive (model ZST25, Thorlabs, Inc.) via a motion control board (KFLOP, DynoMotion). This actuator maintains a constant reciprocating translation of the cuvette along the sample flow direction over a distance of ~ 1 mm at a speed of 0.02 mm s^{-1} . This limits the residence time of the slow-moving protein solution along the walls of the flow cuvette in the laser beam. Scattering from the near-wall areas of the cell was rejected optically. The temperature of gas is maintained at 4 °C by re-heating the nitrogen stream from a dry-ice/ethanol bath with a two-channel temperature controller (Model 34, Cryocon). In addition, the

temperature of the core mixing chamber is further controlled by attached miniature Peltier cells. All wetted surfaces in the TR^3 setup are comprised of protein-compatible PEEK, PTFE (in anaerobic atmosphere), fused silica and glass.

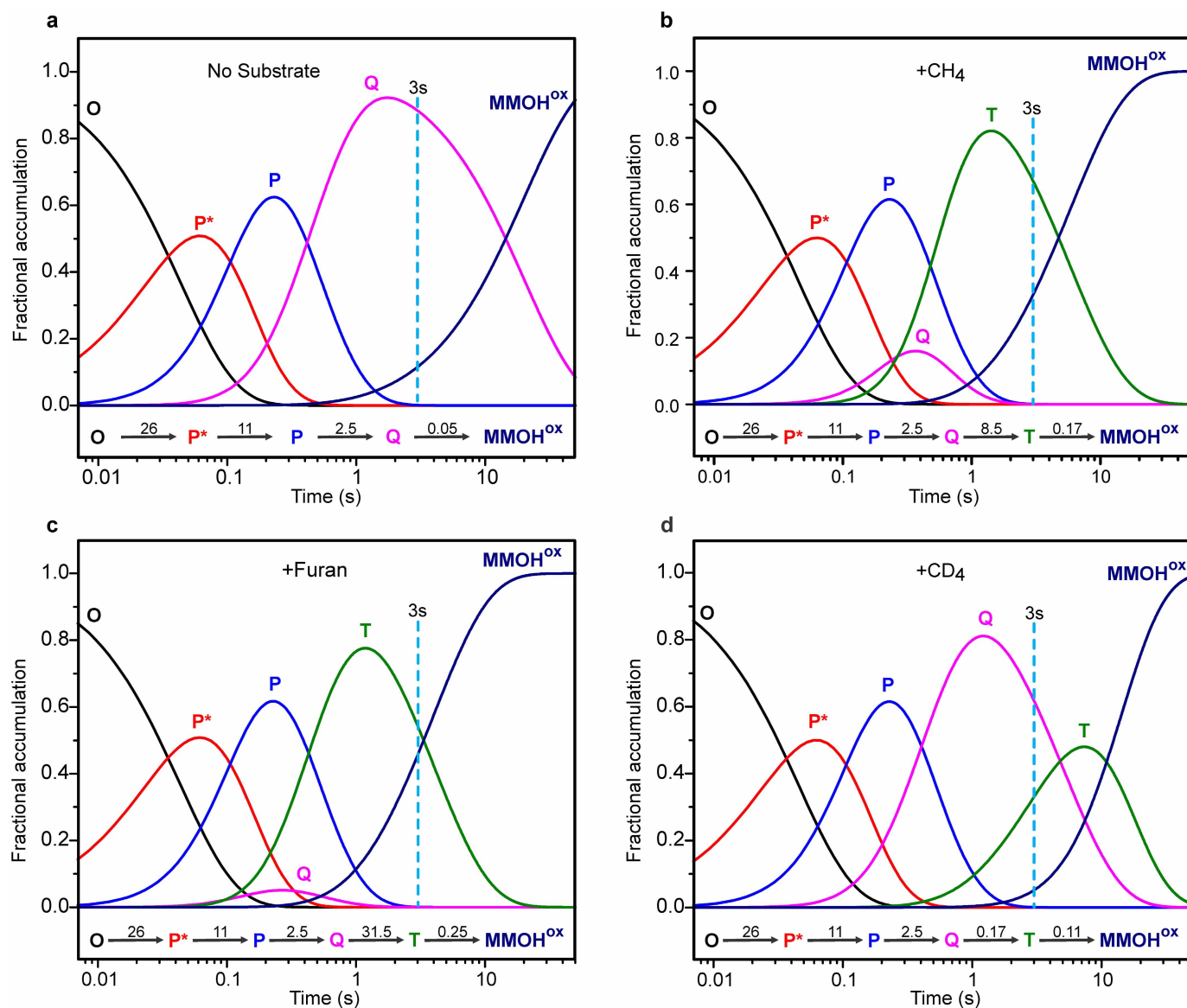
The oxygen-saturated buffer was prepared by addition of 1 ml of O_2 gas to 6 ml of argon-sparged anaerobic buffer at room temperature in a 10 ml gastight syringe. During operation of the TR^3 experiment, the oxygen-saturated buffers were automatically and repeatedly purged and switched from one isotope to the other using computer-controlled actuated HPLC valves (Upchurch Scientific & Rheodyne, IDEX Health & Science). Resonance Raman spectra were collected in 4–12 pairs of oxygen isotope runs in the course of a single TR^3 experiment to randomize the non-isotopic variations in the sample composition. Results of three to eight separate experiments performed over the course of several months using different enzyme preparations were averaged together, taking into account normalization of resonance Raman intensity.

Methane saturated buffer (CH_4 or CD_4) was prepared in an identical fashion to the oxygen saturated buffer. The furan substrate was introduced by adding a calculated volume to the reduced protein sample.

Excitation in the near-ultraviolet (351 nm argon laser line, Coherent, model Innova 300) was used to maximize resonance Raman enhancement and eliminate possibility of interference from porphyrins. The laser power was typically 60 mW and was reduced down to 15 mW during power dependence studies. At a flow rate of 10 μl per min and full laser power an estimated 6,700 photons were absorbed per molecule of Q.

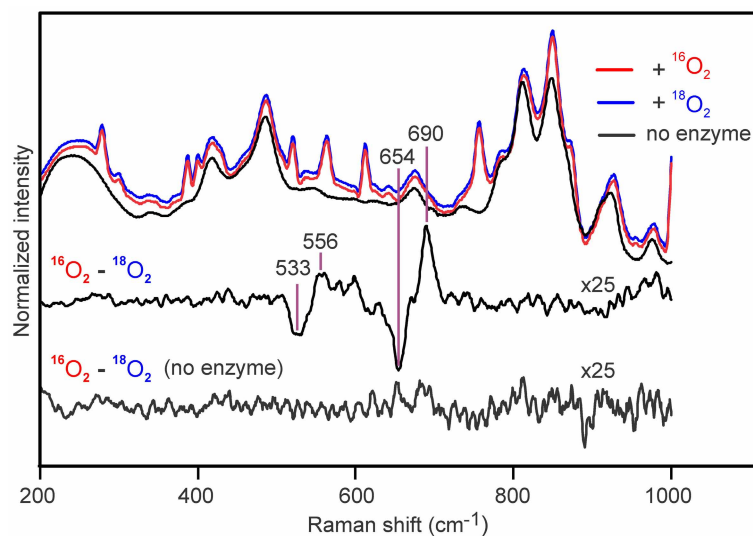
Acetone injected into the flow cell was used as a standard for calibration of the resonance Raman shifts and was subsequently purged with anaerobic buffer. Quantitative analysis of spectral changes over the reaction time or with substrate/solvent substitution was facilitated by normalization to the intensity of a 756 cm^{-1} protein mode as an internal standard. Further details of the spectral analysis are described in the supplementary text of reference¹².

30. Zhang, J. & Lipscomb, J. D. Role of the C-terminal region of the B component of *Methylosinus trichosporium* OB3b methane monooxygenase in the regulation of oxygen activation. *Biochemistry* **45**, 1459–1469 (2006).
31. Fox, B. G., Froland, W. A., Dege, J. E. & Lipscomb, J. D. Methane monooxygenase from *Methylosinus trichosporium* OB3b. Purification and properties of a three-component system with high specific activity from a type II methanotroph. *J. Biol. Chem.* **264**, 10023–10033 (1989).
32. Brazeau, B. J. & Lipscomb, J. D. Kinetics and activation thermodynamics of methane monooxygenase compound Q and reaction with substrates. *Biochemistry* **39**, 13503–13515 (2000).
33. Lee, S. K. & Lipscomb, J. D. Oxygen activation catalyzed by methane monooxygenase hydroxylase component: Proton delivery during the O–O bond cleavage steps. *Biochemistry* **38**, 4423–4432 (1999).
34. Rosenzweig, A. C., Nordlund, P., Takahara, P. M., Frederick, C. A. & Lippard, S. J. Geometry of the soluble methane monooxygenase catalytic diiron center in two oxidation states. *Chem. Biol.* **2**, 409–418 (1995).
35. Srncic, M. *et al.* Structural and spectroscopic properties of the peroxodiferric intermediate of *Ricinus communis* soluble Δ^9 desaturase. *Inorg. Chem.* **51**, 2806–2820 (2012).
36. Jensen, K. P., Bell, C. B. III, Clay, M. D. & Solomon, E. I. Peroxo-type intermediates in class I ribonucleotide reductase and related binuclear non-heme iron enzymes. *J. Am. Chem. Soc.* **131**, 12155–12171 (2009).
37. Han, W. G. & Noodleman, L. Structural model studies for the peroxo intermediate P and the reaction pathway from P→Q of methane monooxygenase using broken-symmetry density functional calculations. *Inorg. Chem.* **47**, 2975–2986 (2008).
38. England, J. *et al.* A synthetic high-spin oxoiron(IV) complex: Generation, spectroscopic characterization, and reactivity. *Angew. Chem. Int. Ed.* **48**, 3622–3626 (2009).
39. Rohde, J.-U. *et al.* Crystallographic and spectroscopic characterization of a nonheme $\text{Fe}(\text{IV})=\text{O}$ complex. *Science* **299**, 1037–1039 (2003).
40. Broadwater, J. A., Ai, J., Loehr, T. M., Sanders-Loehr, J. & Fox, B. G. Peroxodiferric intermediate of stearyl-acyl carrier protein Δ^9 desaturase: Oxidase reactivity during single turnover and implications for the mechanism of desaturation. *Biochemistry* **37**, 14664–14671 (1998).
41. Moenne-Loccoz, P. *et al.* The ferroxidase reaction of ferritin reveals a diferric μ -1,2 bridging peroxide intermediate in common with other O_2 -activating non-heme diiron proteins. *Biochemistry* **38**, 5290–5295 (1999).



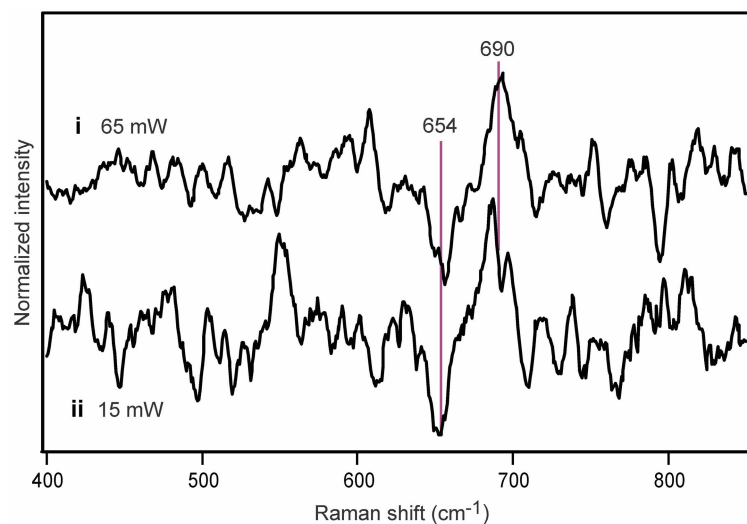
Extended Data Figure 1 | Speciation plots of the sMMO reaction with O₂. Plots were computed using known rate constants of the catalytic steps^{1,8,32,33} for the following conditions. **a**, No added substrate, **b**, presence of 0.45 mM CH₄, **c**, presence of 3.5 mM furan and **d**, presence of 0.45 mM CD₄ at pH 7.0, 4 °C. Rate constants used in simulation of individual conditions are shown for each step. All rate constants are first order (s⁻¹) except for the Q to T step,

which is first order in both Q and substrate (M⁻¹s⁻¹) but is given as a pseudo first order constant for the current substrate concentration. The rate constant for the formation of intermediate O (oxygen binding) is unknown, but it is assumed to be fast based upon typical rates for metalloenzymes of the MMO type. It is irreversible because the rate constant of the next step (O → P*) is independent of O₂ concentration.



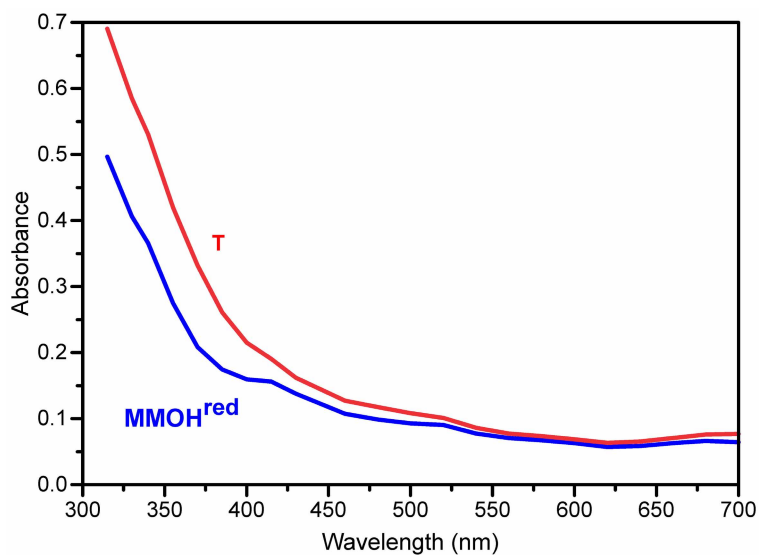
Extended Data Figure 2 | Absolute TR³ spectra of the sMMO reaction with O₂. Top, reaction using ¹⁶O₂-containing buffer (blue), ¹⁸O₂-containing buffer (red) or buffer background in the absence of MMOH/MMOB (black). Middle, ¹⁶O₂ – ¹⁸O₂ difference spectra of the MMOH^{red}/MMOB reaction

with O₂ at pH 7.0, 4 °C and $\Delta t \approx 3.0$ s. Bottom, ¹⁶O₂ – ¹⁸O₂ difference spectra of the oxygenated buffers in the absence of MMOH and MMOB proteins. Intensities of sMMO spectra were normalized to protein vibration. In the absence of protein, relative intensity was normalized using buffer vibrations.



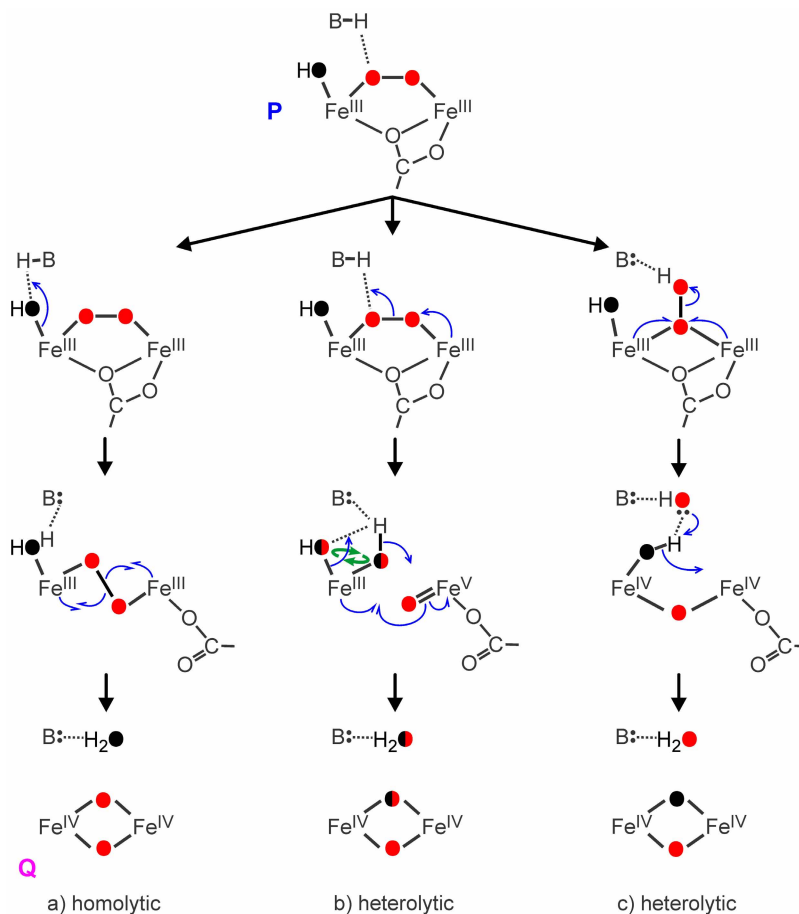
Extended Data Figure 3 | Power dependence of TR³ spectra of sMMO. ¹⁶O₂ – ¹⁸O₂ difference spectra obtained using 65 mW (i) and 15 mW (ii) excitation laser power show the same normalized intensity of oxygen

vibrations in Q, indicating that no detectable photodecomposition is taking place under current conditions.



Extended Data Figure 4 | A comparison of the electronic absorption spectra of compound T (red trace) and MMOH^{red} (blue trace). T exhibits an absorption band in the near-ultraviolet region, giving rise to its resonance Raman enhancement. Single wavelength time courses of the reaction of 25 μM MMOH^{red}/MMOB with a 450 μM solution of CH_4 and 450 μM O_2 at 4 $^\circ\text{C}$,


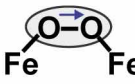

pH 7 were recorded throughout the visible region (concentrations after mixing). The absorbance at each wavelength at the time of maximal T formation given by the speciation plot shown in Extended Data Fig. 1b was extracted and used to make the red trace shown.



Extended Data Figure 5 | Potential O-O bond cleavage mechanisms in the dinuclear centre of MMO. The most divergent mechanisms are shown along with expected isotopic composition of oxygen derived from O_2 (red) and solvent (black). All mechanisms are triggered by proton-dependent rearrangement of **P**^{10,33}. The monodentate carboxylate bridge (E243) found in the diferrous enzyme³⁴ is likely to maintain this position in **P**, but return to the non-bridging position in **Q**, as found in the resting enzyme, to accommodate the diamond core structure. The catalytic base **B**, which mediates proton dependency, has not been definitively identified. Based on structural similarity to other di-iron O_2 -activating enzymes and DFT computations for **P**-analogues in those systems^{11,35,36}, we have proposed¹¹ that E114 (a ligand to solvent-coordinated iron in **P**), is this base. Other ligands not directly involved in cleavage are omitted for clarity (see Fig. 3). Equal intensities of Q - $^{16}O_2$ and Q - $^{18}O_2$ modes, and the absence of Q - $^{16}O^{18}O$ mode in Fig. 2c, (i) argue against isotope scrambling in **Q** formation. This and all other experimental results reported to date are in full accord with the nominally concerted homolytic cleavage mechanism. We postulate that the loss of E243 bridge facilitates the conversion of *cis*- μ -peroxo adduct in **P** to the *trans*- μ -peroxo conformation and the ensuing O-O bond cleavage (a) to form the diamond core structure detected here. This transition is supported by DFT

computations³⁷. In contrast, the stepwise, end-on heterolytic cleavage mechanism (c) (analogous to formation of compound I in cytochrome P450) leads to the mixed isotope cluster in Q - $^{18}O_2$ and can be ruled out. Heterolytic cleavage of *trans*- μ -peroxo is essentially isoelectronic to and experimentally indistinguishable from the homolytic mechanism (a). The proton-assisted heterolytic cleavage of *cis*- μ -peroxo bridge (b) cannot be ruled out yet, but several observations argue against it. (1) We did not observe isotope scrambling (curved green arrows), which is expected upon formation of two terminal oxygenic ligands on the same iron. While scrambling may not occur if ligands are highly stabilized, structural basis for such putative stabilization is not apparent. Scrambling may also not be observed if formation of diamond core is fast following bond cleavage, in which case mechanism (b) becomes, in essence, a stepwise, proton-assisted homolytic cleavage, also indistinguishable from (a). (2) Two iron atoms in di-ferric **P** and di-ferryl **Q** are in the same oxidation states and in indistinguishable electronic environments^{2,20}. Such symmetry is unfavourable for O-O bond polarization and charge separation in the Fe^{III}/Fe^V state during heterolytic cleavage. The deprotonated state of the peroxo bridge in **P** also argues against overall polarity of the site that would aid heterolytic cleavage.

Extended Data Table 1 | Additional vibrational modes of metal centres relevant to Q and T

Vibrational structure	System	$\nu(\Delta^{18}\text{O})$ cm^{-1}	$\nu^{16}\text{O}^{18}\text{O}$ cm^{-1}	$\Delta^2\text{H}$ cm^{-1}	Ref
	sMMO Q	690 (-36)	673	0	This work
	sMMO T	556 (-23)	-	0	This work
	$[\text{Fe}^{\text{IV}}(\text{O})(\text{TMG}_3\text{tren})]$	843 (-33)	-	-	38
	TauD-F4	825 (-37)	-	0	12
	$[\text{Fe}^{\text{IV}}(\text{O})(\text{TMC})(\text{NCCH}_3)]^{2+}$	834 (-34)	-	-	39
	hDOHH _{peroxo}	855 ^a (-44)	833	-	15
	Peroxo- $\Delta^9\text{D}$	898 (-53)	874	-	40
	Frog M ferritin peroxo	851 (-51)	-	-	41
	hDOHH _{peroxo}	473 (-16)	465	-	15
	Peroxo- $\Delta^9\text{D}$	442 (-17)	433	-	40
	Frog M ferritin peroxo	485 (-17)	477	-	41

References 38–41 are cited in this Table. ^a Represents the centre of a Fermi doublet.

Structure of human cytoplasmic dynein-2 primed for its power stroke

Helgo Schmidt^{1*}, Ruta Zalyte^{1*}, Linas Urnavicius¹ & Andrew P. Carter¹

Members of the dynein family, consisting of cytoplasmic and axonemal isoforms, are motors that move towards the minus ends of microtubules. Cytoplasmic dynein-1 (dynein-1) plays roles in mitosis and cellular cargo transport¹, and is implicated in viral infections² and neurodegenerative diseases³. Cytoplasmic dynein-2 (dynein-2) performs intraflagellar transport⁴ and is associated with human skeletal ciliopathies⁵. Dyneins share a conserved motor domain that couples cycles of ATP hydrolysis with conformational changes to produce movement^{6–9}. Here we present the crystal structure of the human cytoplasmic dynein-2 motor bound to the ATP-hydrolysis transition state analogue ADP.vanadate¹⁰. The structure reveals a closure of the motor's ring of six AAA+ domains (ATPases associated with various cellular activities: AAA1–AAA6). This induces a steric clash with the linker, the key element for the generation of movement, driving it into a conformation that is primed to produce force. Ring closure also changes the interface between the stalk and buttress coiled-coil extensions of the motor domain. This drives helix sliding in the stalk which causes the microtubule binding domain at its tip to release from the microtubule. Our structure answers the key questions of how ATP hydrolysis leads to linker remodelling and microtubule affinity regulation.

There are four nucleotide-binding sites in the dynein motor, but movement only depends on ATP hydrolysis in the first site (AAA1)^{7,11,12}. When this site is nucleotide free or bound to ADP, the microtubule binding domain (MTBD) binds to the microtubule and the linker adopts the straight post-power-stroke conformation^{6–8,12–14}. Upon ATP binding and hydrolysis, the MTBD detaches from the microtubule and the linker is primed into the pre-power-stroke conformation^{6,12,14,15} (Fig. 1a). MTBD rebinding leads to a force producing swing of the linker (power stroke) back to the post-power-stroke position and the release of ATP hydrolysis products to reset the cycle^{6,14–16}.

To address how the linker is primed and dynein released from microtubules, we co-crystallized the human dynein-2 motor domain with ADP.vanadate (ADP.Vi) to trap it in a pre-power-stroke state⁶ (Extended Data Fig. 1 and Extended Data Table 1). The linker in this dynein-2:ADP.Vi structure has a 90° bend (Fig. 1b) consistent with low-resolution studies of pre-power-stroke dynein^{6,8,9,17}. Dynein's AAA+ domains are each divided into an α/β 'large' subdomain (AAAL, helices H0–H4 and beta strands S1–S5) and an α 'small' subdomain (AAAS, helices H5–H9)¹⁶. The individual subdomains of dynein-2:ADP.Vi are highly similar to those in post-power-stroke crystal structures of dynein-1 from *Saccharomyces cerevisiae*¹³ (dynein-1:APO; Protein Data Bank (PDB) accession number 4AKI) and *Dictyostelium discoideum*¹⁸ (dynein-1:ADP; PDB accession number 3VKG) (Extended Data Fig. 2 and Supplementary Data 1). This suggests conformational changes between these structures (Supplementary Discussion and Extended Data Fig. 3a) are not related to sequence differences but are caused by the different nucleotide states.

In dynein-2:ADP.Vi, all four nucleotide-binding sites are occupied (Fig. 1c). The AAA1 site, found between AAA+ domains AAA1 and AAA2, binds ADP.Vi (Fig. 2a and Extended Data Fig. 4a–d) via conserved motifs¹⁹ (Fig. 2b). The trigonal-bipyramidal vanadate group mimics the

ATP γ -phosphate during hydrolysis¹⁰. It is surrounded by three important catalytic residues¹⁹: the Walker B glutamate (W-B: E1742), the sensor-I asparagine (S-I: N1792) and the AAA2L arginine finger (RF: R2109), suggesting the structure is in the ATP hydrolysis-competent conformation.

In the dynein-1 structures there is a gap between AAA1 and AAA2. The closure of this gap in dynein-2:ADP.Vi (Supplementary Video 1) is driven by the arginine finger-ADP.Vi contact. It is reinforced by additional interactions between AAA1L and AAA2L (Fig. 2c). A pair of conserved inserts in AAA2L²⁰ (the 'H2 insert' and the 'pre-sensor-I' (PS-I) insert) contact the H2 helix in AAA1L (Fig. 2a, c) and displace H2 and H3 relative to the rest of AAA1L (Extended Data Fig. 5a). The AAA1L sensor-I loop, which varies in position depending on dynein's nucleotide state (Extended Data Fig. 5b), swings in to contact AAA2L (Fig. 2c).

The other nucleotide-binding sites contain tightly bound nucleotides that co-purify with the motor domain (Extended Data Fig. 4e–j and Supplementary Discussion). The density in AAA2 is consistent with an ATP, as observed in all the dynein structures^{13,18}. As in dynein-1:ADP, the densities in AAA3 and AAA4 suggest the presence of ADP. In the dynein-1:APO structure these sites are empty. In all dynein structures

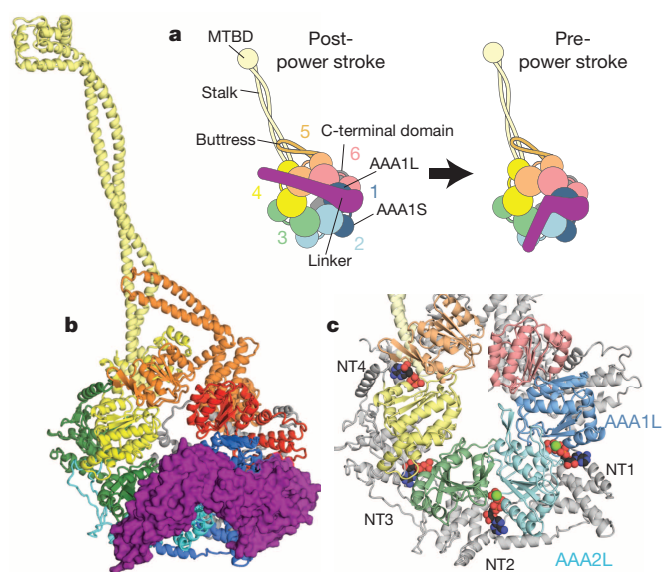


Figure 1 | Crystal structure of dynein-2:ADP.Vi. **a**, Schematic representation of a dynein motor domain in post- and pre-power-stroke states. Structural elements are labelled and colour-coded. AAA+ domains (1–6) consist of large (AAAL) and small (AAAS) subdomains. The coiled-coil stalk is supported by the coiled-coil buttress and harbours the MTBD. A carboxy (C)-terminal domain runs underneath the AAA+ ring. **b**, Overview of dynein-2:ADP.Vi in cartoon/surface representation. The linker features a 90° bend. **c**, Nucleotides (NT1–NT4, sphere representations) are mainly bound between AAA+ large domains (colour-coded). AAA1L and AAA2L form the important AAA1 nucleotide-binding site.

¹Medical Research Council Laboratory of Molecular Biology, Division of Structural Studies, Francis Crick Avenue, Cambridge CB2 0QH, UK.

*These authors contributed equally to this work.

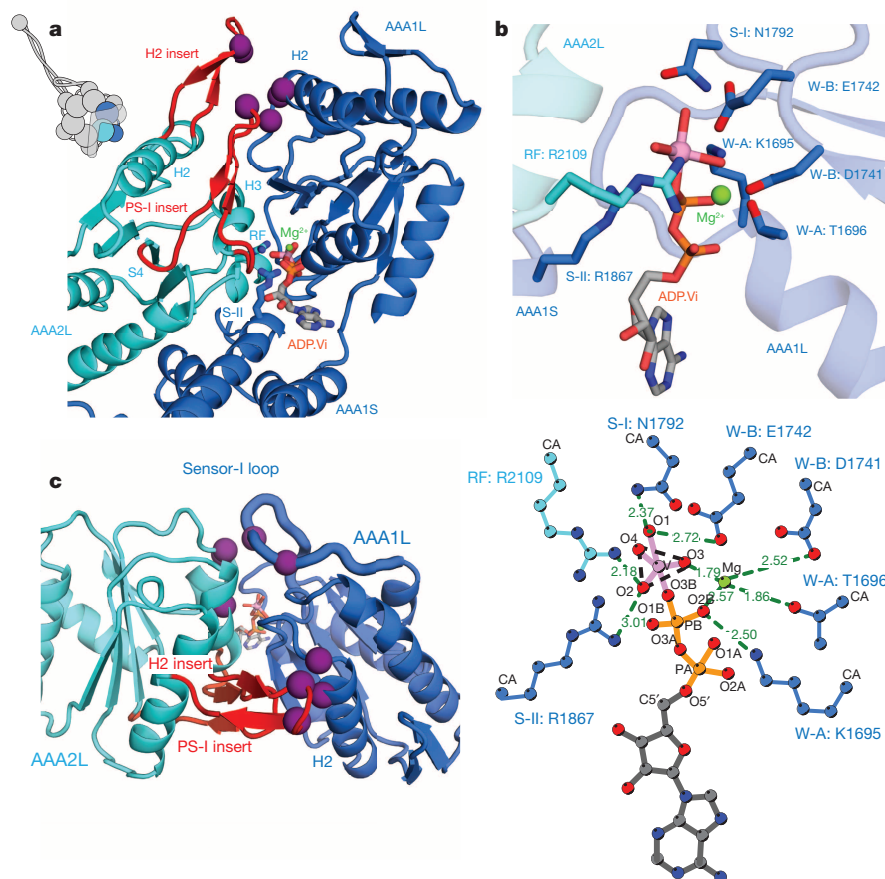


Figure 2 | ADP.Vi binding to AAA1 nucleotide-binding site induces closure of AAA1/AAA2 interface. **a**, AAA1L, AAA1S and AAA2L enclose ADP.Vi. The AAA2L H2 and PS-I inserts (red) contact AAA1L H2. **b**, Upper panel: the Mg²⁺.ADP contacts the Walker A (W-A: K1695), Walker B (W-B: D1741) and the sensor-II (S-II: R1867) residues. The trigonal-bipyramidal Vi-group mimics the ATP-hydrolysis transition state and is surrounded by sensor-I

(S-I: N1792) and Walker B (W-B: E1742) residues and the AAA2L arginine finger (RF). Lower panel: schematic diagram showing the distances from ADP.Vi to the catalytic residues (in Å). **c**, AAA1/AAA2 interface closure is reinforced by the AAA1L sensor-I loop contacting AAA2L. Purple spheres represent contacts.

the AAA2 and AAA3 nucleotide-binding sites are in a similar closed conformation. This means that the whole AAA2–AAA4 region forms a rigid block (Extended Data Fig. 5c, d).

The linker, which is divided into four subdomains^{13,18}, bends between subdomains 2 and 3 in the pre-power-stroke dynein-2:ADP.Vi structure. Compared with the straight post-power-stroke linker, the mobile subdomains 1 and 2 (Link1–2, helices H5–H9) undergo a rigid-body movement relative to the static subdomains 3 and 4 (Link3–4, helices H11–H18) (Fig. 3a and Supplementary Video 2). The hinge helix (H10), which connects Link1–2 and Link3–4, is forced to adopt a curved conformation. The isolated linker prefers a straight conformation⁸, suggesting the distorted hinge helix is strained and can act as a store of energy.

In all dynein structures, the static Link3–4 is connected to the AAA+ ring via contacts to AAA1 (Extended Data Fig. 6a–c). The closure of the AAA1 site in dynein-2:ADP.Vi establishes two additional interactions (Fig. 3b). The AAA2L PS-I insert contacts the loop between H11 and H12 on Link3 via a backbone interaction. The displacement of AAA1L H2 described above allows arginine R1726 to contact Link3 via glutamate E1420.

The mobile Link1–2 region interacts differently with the AAA+ ring in all dynein structures^{13,18} (Extended Data Fig. 6d–f). In dynein-2:ADP.Vi its position is stabilized by conserved hydrophobic interactions across the linker bend (Fig. 3c and Extended Data Fig. 7). The Link1–2 region also makes two interactions with the AAA+ ring (Fig. 3d). One minor contact involves a residue (E2028) on the AAA2L H2-insert. The other, with the AAA3L H2–S3 insert, is more extensive but involves poorly conserved residues (Extended Data Fig. 7).

We had anticipated that the movement of the mobile part of the linker would be driven by its interaction with the highly conserved inserts in AAA2L^{13,14,18}. It was therefore surprising to find that these inserts contact only the static part of the linker. How then could AAA1 site closure induce linker bending? To address this question we asked what would happen if the AAA+ ring adopted the ADP.Vi state but the linker remained in the straight, post-power-stroke conformation. The rigid-body behaviour of AAA2–AAA4 means that closure of the AAA1 site would lead to a steric clash between the mobile Link1–2 region and the AAA4L PS-I insert (Fig. 3e). This is demonstrated by the overlap between these regions observed in an alignment of the straight linker from dynein-1:ADP onto the dynein-2:ADP.Vi structure (Fig. 3f). The additional contacts between the AAA+ ring and the static Link3–4 (Fig. 3b) prevent it moving and mean that the clash can only be relieved by the mobile Link1–2 adopting its pre-power-stroke position (Fig. 3e, g and Supplementary Video 3).

To test this model we used negative-stain electron microscopy to assay linker movement (Fig. 3h, i and Extended Data Fig. 8a, b). In the presence of ADP all dynein-2 motors had an angle between the stalk and linker of $54 \pm 13^\circ$ (mean \pm s.d.) (Fig. 3h). In the presence of ADP.Vi the majority of motors showed a pre-power-stroke conformation with an angle of $145 \pm 20^\circ$. We then tested the ability of dynein-2 mutants to adopt the pre-power-stroke state in the presence of ADP.Vi (Fig. 3i). Removal of the AAA2L inserts abolished the linker movement, consistent with previous data¹⁸. It also completely prohibited microtubule gliding activity (Extended Data Fig. 8c). When the AAA4L PS-I insert was deleted only a small percentage of motors attained the pre-power-stroke

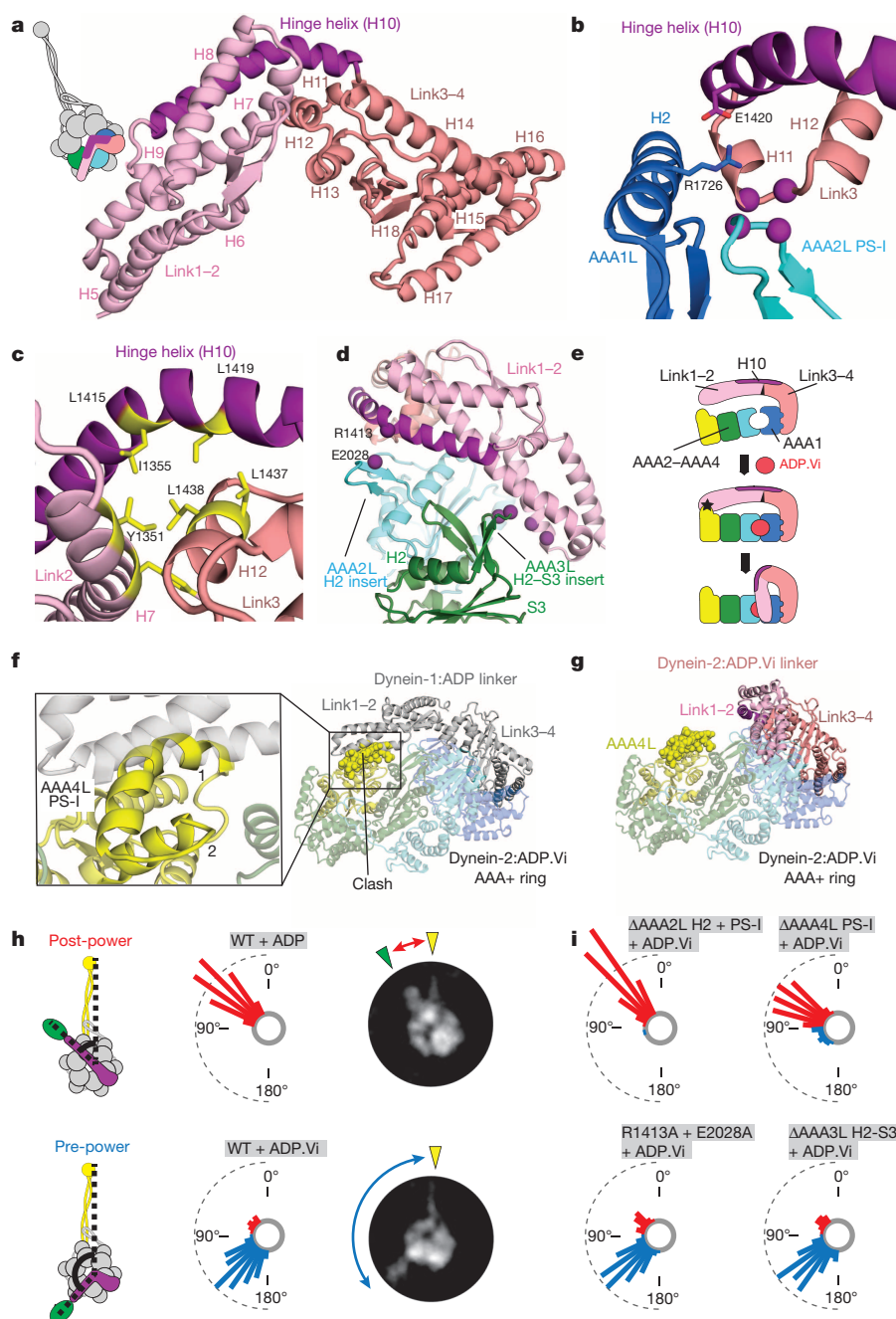


Figure 3 | Linker bending upon closure of AAA1 nucleotide-binding site.

a, A 90° bend between linker subdomains 1 and 2 (Link1–2) and 3 and 4 (Link3–4) forces the hinge helix (H10) to curve. **b**, The AAA2L PS-I insert contacts Link3 and AAA1L R1726 forms a salt bridge with E1420 on the hinge helix. **c**, Hydrophobic residues (yellow) stabilize the Link2/Link3 bend. **d**, Link1–2 contacts the AAA2:H2 and AAA3L:H2–S3 inserts. **e**, AAA1 site (blue/cyan) closure causes a rigid-body movement of AAA2–AAA3–AAA4 (cyan–green–yellow) leading to a clash (black star) with Link1–2 (light pink). To relieve the clash, the linker adopts the pre-power-stroke conformation. **f**, The straight post-power-stroke linker (grey), aligned via Link3–4 onto dynein-2:ADP.Vi, would clash with the AAA4L PS-I insert (yellow spheres). **g**, In dynein-2:ADP.Vi the linker moved to avoid the clash. **h**, In a

negative-stain electron microscopy assay, the angle between the stalk (yellow) and green fluorescent protein (GFP)–linker (green/purple) of the dynein-2:ADP motor is $54 \pm 13^\circ$ (mean \pm s.d.). With ADP.Vi most dynein-2 motors are in a pre-power-stroke state with an angle of $145 \pm 20^\circ$ (mean \pm s.d.). **i**, Deletion of the AAA2L (Δ AAA2L:H2 + PS-I) or AAA4L inserts (Δ AAA4L:PS-I) hinders the linker adopting the pre-power-stroke conformation with ADP.Vi. Mutation of either AAA3 (Δ AAA3L:H2–S3) or AAA2 (R1413A + E2028A) linker-ring contacts has no effect. Dashed half-circles in **h** and **i** mark 20% dynein-2 particles. All negative-stain electron microscopy experiments were done in triplicate. The number of particles used in each experiment are provided in Methods.

conformation (Fig. 3i), supporting our model that the AAA4L PS-I insert plays a major role in linker bending. In agreement with this interpretation, the microtubule gliding velocity of this mutant was only 10% of wild type (Extended Data Fig. 8c). In contrast, removal of the Link1–2 contacts with AAA2L and AAA3L had a minimal effect on linker movement (Fig. 3i).

In addition to triggering movement of the linker, ADP.Vi binding to dynein reduces the affinity of its MTBD for microtubules²¹. Biochemical^{22,23} and structural^{14,18,24–26} evidence suggests this involves the helices in the stalk, coiled-coil helix 1 (CC1) and coiled-coil helix 2 (CC2), sliding past each other by one turn of α -helix. The dynein-2:ADP.Vi structure, where the MTBD has low microtubule affinity (Extended Data Fig. 9

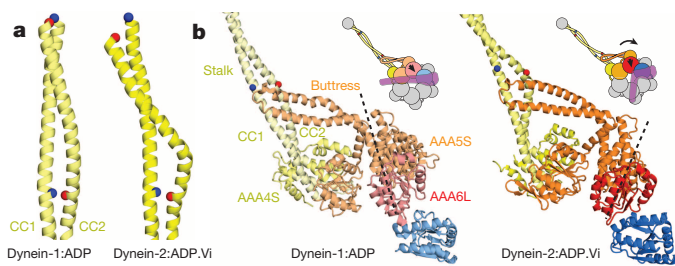


Figure 4 | Butress movement triggers helix sliding in the stalk. **a**, In dynein-2:ADP.Vi the stalk CC2 is kinked. Compared with the dynein-1:ADP stalk, CC2 is displaced by one turn of α -helix relative to CC1. Blue and red spheres represent equivalent amino-acid residues in the two structures. **b**, The two stalk conformations are stabilized by movement of the buttress, which slips relative to CC1, but moves with CC2 when comparing the two structures. The buttress movement is the result of a rotation of the AAA6L/AAA5S unit (shown by the position of the dashed black line).

and Supplementary Discussion), answers the key question of how the sliding is initiated.

In dynein-2:ADP.Vi the base of the stalk deviates from the symmetrical, regular coiled coil observed in dynein-1:ADP (Fig. 4a). The stalk CC2 helix contains a kink, located near the stalk/buttress interface, which causes it to slip relative to CC1. The resulting asymmetry between the two helices is similar to that observed in the parallel coiled-coil homodimer Bicaudal-D²⁷. A comparison of the dynein-1:ADP and dynein-2:ADP.Vi structures (Fig. 4b) suggests how the movement of the buttress, relative to the stalk, is coupled to the movement of CC2. In dynein-2:ADP.Vi the buttress slides relative to CC1 but moves together with CC2.

The stalk and buttress emerge from AAA4S and AAA5S respectively. Their relative movement is coupled to rearrangements in the AAA+ ring. Closure of the AAA1 site and the rigid body movement of AAA2–AAA4 force the AAA4/AAA5 interface to close and the AAA6L subdomain to rotate towards the ring centre (Fig. 4b and Supplementary Discussion). The AAA5S subdomain rotates as a unit together with AAA6L, and this movement pulls the buttress relative to the stalk (Supplementary Videos 4 and 5).

Unlike myosin and kinesin motors, dynein shares mechanistic similarities with AAA+ proteins that remodel their substrates²⁸. In dynein, one substrate is the linker which is bent by a clash with the AAA+ ring. This bent conformation is stabilized by contacts at the Link2/Link3 interface, the importance of which is highlighted by the fact that a mutation there (G1442D) can cause the human ciliopathy Jeune syndrome⁵ (Supplementary Discussion). When the AAA1 site reopens, the bent linker reverts to its preferred straight conformation⁸ and generates force. In addition to the linker, the dynein AAA+ ring also remodels the stalk. Here the motions of AAA+ domains are directly coupled to sliding of helices in the coiled coil (Supplementary Video 6).

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 11 August; accepted 29 October 2014.

Published online 1 December 2014.

1. Roberts, A. J., Kon, T., Knight, P. J., Sutoh, K. & Burgess, S. A. Functions and mechanics of dynein motor proteins. *Nature Rev. Mol. Cell Biol.* **14**, 713–726 (2013).
2. Dodding, M. P. & Way, M. Coupling viruses to dynein and kinesin-1. *EMBO J.* **30**, 3527–3539 (2011).
3. Schiavo, G., Greensmith, L., Hafezparast, M. & Fisher, E. M. Cytoplasmic dynein heavy chain: the servant of many masters. *Trends Neurosci.* **36**, 641–651 (2013).
4. Ishikawa, H. & Marshall, W. F. Ciliogenesis: building the cell's antenna. *Nature Rev. Mol. Cell Biol.* **12**, 222–234 (2011).

5. Schmidts, M. *et al.* Exome sequencing identifies DYNC2H1 mutations as a common cause of asphyxiating thoracic dystrophy (Jeune syndrome) without major polydactyly, renal or retinal involvement. *J. Med. Genet.* **50**, 309–323 (2013).
6. Burgess, S. A., Walker, M. L., Sakakibara, H., Knight, P. J. & Oiwa, K. Dynein structure and power stroke. *Nature* **421**, 715–718 (2003).
7. Kon, T., Mogami, T., Ohkura, R., Nishiura, M. & Sutoh, K. ATP hydrolysis cycle-dependent tail motions in cytoplasmic dynein. *Nature Struct. Mol. Biol.* **12**, 513–519 (2005).
8. Roberts, A. J. *et al.* ATP-driven remodeling of the linker domain in the dynein motor. *Structure* **20**, 1670–1680 (2012).
9. Roberts, A. J. *et al.* AAA+ ring and linker swing mechanism in the dynein motor. *Cell* **136**, 485–495 (2009).
10. Davies, D. R. & Hol, W. G. The power of vanadate in crystallographic investigations of phosphoryl transfer enzymes. *FEBS Lett.* **577**, 315–321 (2004).
11. Gibbons, I. R., Gibbons, B. H., Moczy, G. & Asai, D. J. Multiple nucleotide-binding sites in the sequence of dynein β heavy chain. *Nature* **352**, 640–643 (1991).
12. Kon, T., Nishiura, M., Ohkura, R., Toyoshima, Y. Y. & Sutoh, K. Distinct functions of nucleotide-binding/hydrolysis sites in the four AAA modules of cytoplasmic dynein. *Biochemistry* **43**, 11266–11274 (2004).
13. Schmidt, H., Gleave, E. S. & Carter, A. P. Insights into dynein motor domain function from a 3.3-Å crystal structure. *Nature Struct. Mol. Biol.* **19**, 492–497 (2012).
14. Carter, A. P. Crystal clear insights into how the dynein motor moves. *J. Cell Sci.* **126**, 705–713 (2013).
15. Imamura, K., Kon, T., Ohkura, R. & Sutoh, K. The coordination of cyclic microtubule association/dissociation and tail swing of cytoplasmic dynein. *Proc. Natl Acad. Sci. USA* **104**, 16134–16139 (2007).
16. Carter, A. P., Cho, C., Jin, L. & Vale, R. D. Crystal structure of the dynein motor domain. *Science* **331**, 1159–1165 (2011).
17. Lin, J., Okada, K., Raychev, M., Smith, M. C. & Nicastro, D. Structural mechanism of the dynein power stroke. *Nature Cell Biol.* **16**, 479–485 (2014).
18. Kon, T. *et al.* The 2.8-Å crystal structure of the dynein motor domain. *Nature* **484**, 345–350 (2012).
19. Wendler, P., Ciniawsky, S., Kock, M. & Kube, S. Structure and function of the AAA+ nucleotide binding pocket. *Biochim. Biophys. Acta* **1823**, 2–14 (2012).
20. Gleave, E. S., Schmidt, H. & Carter, A. P. A structural analysis of the AAA+ domains in *Saccharomyces cerevisiae* cytoplasmic dynein. *J. Struct. Biol.* **186**, 367–375 (2014).
21. Vale, R. D., Soll, D. R. & Gibbons, I. R. One-dimensional diffusion of microtubules bound to flagellar dynein. *Cell* **59**, 915–925 (1989).
22. Gibbons, I. R. *et al.* The affinity of the dynein microtubule-binding domain is modulated by the conformation of its coiled-coil stalk. *J. Biol. Chem.* **280**, 23960–23965 (2005).
23. Kon, T. *et al.* Helix sliding in the stalk coiled coil of dynein couples ATPase and microtubule binding. *Nature Struct. Mol. Biol.* **16**, 325–333 (2009).
24. Carter, A. P. *et al.* Structure and functional role of dynein's microtubule-binding domain. *Science* **322**, 1691–1695 (2008).
25. Nishikawa, Y. *et al.* Structure of the entire stalk region of the dynein motor domain. *J. Mol. Biol.* **426**, 3232–3245 (2014).
26. Redwine, W. B. *et al.* Structural basis for microtubule binding and release by dynein. *Science* **337**, 1532–1536 (2012).
27. Liu, Y. *et al.* Bicaudal-D uses a parallel, homodimeric coiled coil with heterotypic registry to coordinate recruitment of cargos to dynein. *Genes Dev.* **27**, 1233–1246 (2013).
28. Erzberger, J. P. & Berger, J. M. Evolutionary relationships and structural mechanisms of AAA+ proteins. *Annu. Rev. Biophys. Biomol. Struct.* **35**, 93–114 (2006).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was funded by the Medical Research Council, UK (MC_UP_A025_1011), a Wellcome Trust New Investigator Award (WT100387) and an EMBO Young Investigator Award. We thank M. Yu for in-house support with X-ray data collection and Diamond Light Source for access to beamline I02 (MX8547-70). We thank G. Dornan and M. Barczyk for assistance with insect cell culture. We also thank S. Bullock and D. Barford for their advice and comments on the manuscript.

Author Contributions R.Z. and H.S. screened many dynein species for expression and crystallization. R.Z. expressed human dynein-2 in insect cells, obtained crystals in the presence of vanadate and collected data. H.S. phased the structure and built an initial model. A.P.C. built and refined the structure. R.Z. and H.S. made mutants and performed biochemical assays. L.U. performed negative-stain electron microscopy. H.S., R.Z. and A.P.C. prepared the manuscript.

Author Information Coordinates and structure factors have been deposited in the Protein Data Bank under accession number 4RHH. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.P.C. (cartera@mrc-lmb.cam.ac.uk).

METHODS

Cloning of constructs. DNA sequence coding for a variant of human cytoplasmic dynein-2 isoform 1 (GenBank reference number BAG06721), codon-optimized for expression in *Spodoptera frugiperda* (Sf9) cells, was amplified (coding region D1091–Q4307) using Phusion High-Fidelity DNA Polymerase (New England Biolabs). The primers used for construct amplification contained sites homologous to a pFastBac vector (Invitrogen Life Science Technologies) that had been modified to contain a tobacco etch virus protease cleavable tandem Protein A-tag for purification followed by a GFP. InFusion (Clontech Laboratories) was used to insert the *dynein-2_{D1091–Q4307}* gene into the pFastBac vector. The final construct used for crystallization, electron microscopy and microtubule gliding assays had an amino (N)-terminal GFP, followed by a glycine (G) serine (S) spacer and dynein-2_{D1091–Q4307} (GFP–dynein-2_{D1091–Q4307}). All mutants were prepared by standard cloning techniques using GFP–dynein-2_{D1091–Q4307} in the pFastBac vector background as a template. ΔAAA2L H2 + PS-I had regions 2022–2030 and 2074–2085 replaced by GG, ΔAAA4L PS-I had region 2734–2774 replaced by GSGSG, ΔAAA3L H2–S3 had region 2339–2344 replaced by GG and R1413A + E2028A had R1413 and E2028 substituted by alanines. All constructs were sequence verified.

Protein expression in Sf9 cells. The modified pFastBac plasmids were transformed into a DH10 EMBAcy *Escherichia coli* strain which carried a bacmid harbouring the baculovirus genome. Clones containing bacmids in which the pFastBac vector had been successfully integrated were selected by blue white screening. Recombinant bacmids were prepared according to standard procedures, transfected into 2 ml Sf9 cells (0.5×10^6 cells per millilitre) using FuGENE HD Transfection Reagent (Promega) and incubated at 27 °C for 72 h (P1 virus). Half a millilitre of P1 virus was subsequently used to infect 50 ml of Sf9 cells (2×10^6 cells per millilitre) followed by incubation at 27 °C and 127 r.p.m. for 72 h (P2 virus). Five millilitres of P2 virus were used to infect 500 ml of Sf9 cells (2×10^6 cells per millilitre) followed by the incubation procedure described before. Cells were harvested by centrifugation at 4 °C and 2,500g for 30 min. The pellet was washed in ice-cold PBS, snap-frozen in liquid nitrogen and stored at –80 °C.

Protein purification. Frozen pellets were resuspended in lysis buffer (30 mM HEPES pH 7.4, 50 mM KOAc, 2 mM MgOAc, 0.2 mM EGTA, 10% v/v glycerol, 300 mM KCl, 0.2 mM Mg.ATP, 1 mM DTT and 2 mM PMSF). Resuspended cells were lysed manually in a dounce homogenizer. Cell debris and insoluble proteins were removed by ultracentrifugation at 4 °C and 60,000g for 30 min. Dynein constructs were pulled out from the lysate using IgG sepharose beads (GE Healthcare, 5 ml of beads per litre of Sf9 culture). IgG sepharose beads were washed with 15 bead volumes of lysis and tobacco etch virus protease buffer (50 mM Tris HCl pH 8, 150 mM KOAc, 2 mM MgOAc, 1 mM EGTA, 10% v/v glycerol, 1 mM DTT and 0.2 mM Mg.ATP). Protein was released from the beads during overnight cleavage with tobacco etch virus protease. Size-exclusion chromatography was performed on a Superose 6 column (GE Healthcare) in size-exclusion chromatography buffer (20 mM Tris HCl pH 8.0, 100 mM KOAc, 2 mM MgOAc, 1 mM EGTA, 10% v/v glycerol, 1 mM DTT).

Protein crystallization. Peak fractions of GFP–dynein-2_{D1091–Q4307} after size-exclusion chromatography were pooled and concentrated to 8 mg ml^{–1}. To lock dynein in its pre-power-stroke state, Mg.ATP (Sigma Aldrich) and Na₃VO₄ (New England Biolabs) were added to a final concentration of 3 mM each. Crystals were obtained by hanging-drop vapour diffusion at 19 °C, mixing equal volumes of protein with reservoir solution (4–6% PEG 6000 and 0.1 M Tris pH 8.0). Crystallization strictly depended on the presence of both Mg.ATP and Na₃VO₄. Crystals did not form under apo, Mg.ATP or Na₃VO₄ conditions. The crystal quality was markedly improved by microseeding. Seeds were prepared by harvesting GFP–dynein-2_{D1091–Q4307} crystals into 100 μl of reservoir solution followed by vortexing with a seed bead (Jena Bioscience) for 30 s. After diluting the seed stock 1:10,000, crystallization was performed by mixing equal volumes of protein and seeds in reservoir solution, followed by equilibration against reservoir as described above.

Data collection and structure determination. Owing to the very fragile nature of the crystals, cryoprotection was performed in the drop by adding 1 drop volume of reservoir solution supplemented with 60% PEG 400. After incubation for 10–30 s, crystals were harvested using MicroMeshes (MiteGen) and flash cooled in liquid nitrogen. Heavy atom derivatization was performed by adding solid Na₃[PW₁₂O₄₀] × H₂O (Tri-Sodium phosphotungstate, Jena Bioscience) directly to the drop. After incubation for 2 h, crystals were collected as described above. Diffraction data were collected at 100 K on beamline I02 at the Diamond Light Source. The data were integrated using MOSFLM²⁹ and scaled using AIMLESS³⁰. In the case of the anisotropic Native-1 data set, a first round of integration and scaling was performed with a resolution limit of 3 Å. The data were then subsequently analysed with the University of California, Los Angeles Molecular Biology Diffraction Anisotropy Server³¹ (<http://services.mbi.ucla.edu/anisotropy/>), which suggested resolution cut-offs of $a = 4.2$ Å, $b = 4.4$ Å and $c = 3.4$ Å. The second round of data integration and scaling was done with the resolution cut-offs mentioned above (Extended Data Table 1). Phasing was performed with AUTOSHARP³² using the multiple isomorphous

replacement with anomalous scattering (MIRAS) approach with a low-resolution Native-2 data set and the Na₃[PW₁₂O₄₀] peak and inflection data sets as heavy atom derivatives (Extended Data Table 1). The final electron-density map after density modification was of sufficient quality to identify the location of all dynein-2 subdomains in the asymmetric unit. Homology models for the individual subdomains were obtained by combining PDB accession number 3VKG with the amino-acid sequence of human cytoplasmic dynein-2 isoform 1 using the PHYRE server³³. The homology models were placed in the asymmetric unit followed by iterative rounds of refinement in REFMAC³⁴ against the Native-1 data set, employing the 'Jelly-Body' and 'secondary structure restraints' refinement options, and manual rebuilding in COOT³⁵. Refining the model against the anisotropy-corrected data obtained from the University of California, Los Angeles Molecular Biology Diffraction Anisotropy Server³¹ significantly improved the quality of the resulting electron-density maps. The final model was evaluated by calculating a simulated annealing composite omit map in CNS³⁶ and had 99.9% of the residues in the allowed regions of the Ramachandran plot. All figures were prepared using PYMOL (<http://www.pymol.org>), LIGPLOT³⁷ and Jalview³⁸.

Vanadate-mediated ultraviolet-photo cleavage of dynein-2 crystals. Crystals obtained under the conditions described above were harvested and washed three times in 10 μl reservoir solution followed by exposure to ultraviolet light (254 nm) for 1 h. Crystals were subsequently dissolved in sample buffer, boiled for 10 min at 95 °C and analysed by SDS–polyacrylamide gel electrophoresis.

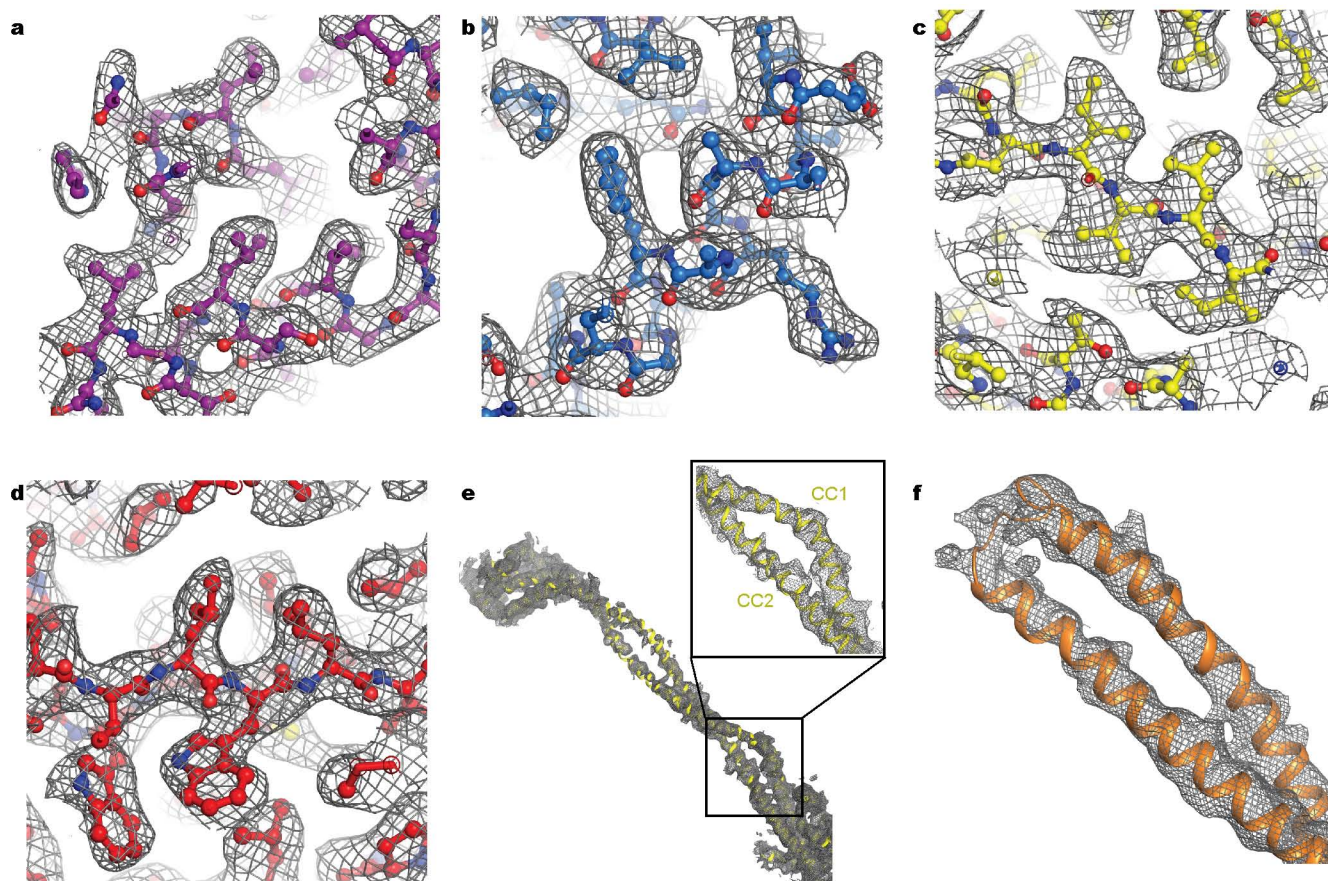
Nucleotide content analysis. Nucleotides were extracted from dynein-2 samples that had been purified as described above. Nucleotide extraction was performed essentially as described previously³⁹. Briefly, concentrated protein was precipitated by adding HClO₄ to a final concentration of 0.5 M. The sample was vortexed and centrifuged for 10 min at 4 °C and 14,000g. The supernatant was mixed with 1 M K₂HPO₄, 3 M KOH and 100% acetic acid (final concentrations were 125 mM, 375 mM and 0.5 M respectively). Total nucleotide content and protein concentration before extraction were measured using a NanoDrop ND-1000 spectrophotometer.

Negative-stain electron microscopy. Dynein motor domain constructs were purified as described above and diluted into electron microscopy assay buffer (50 mM Tris HCl pH 8, 150 mM KOAc, 2 mM MgOAc, 1 mM EGTA and 0.1 mM DTT) to a final concentration of 30 nM. Negative-stain electron microscopy was performed either in the presence of 3 mM Mg.ATP (+ADP) or 3 mM Mg.ATP and 3 mM Na₃VO₄ (+ADP.Vi) on plasma-cleaned carbon film on 400-square-mesh copper grids (Electron Microscopy Sciences). The samples were stained with 2% (w/v) uranyl acetate. Electron micrographs (Extended Data Fig. 8a) were recorded on a Gatan Orius SC200B CCD (charge-coupled device) fitted to a FEI Tecnai G2 Spirit transmission electron microscope operating at 120 kV. Data were collected at ~1 μm underfocus, with a pixel size of 3.29 Å and an estimated dose of 20 electrons per square ångström during 1 s exposures. Automated particle picking was done in EMAN2.10a⁴⁰ using the Swarm boxing tool. Subsequent particle analysis used RELION⁴¹. Autopicked particles were subjected to two-dimensional classification to identify incorrectly picked particles, which were manually checked and removed from the data set. The remaining particles were classified into ten classes, which was sufficient to represent all observed views. Each class was then subclassified into 50 subclasses. Noisy subclasses were discarded and those remaining contained sufficient signal to noise to identify the stalk and linker–GFP clearly (Extended Data Fig. 8b). The ImageJ azimuthal average plugin (<http://rsb.info.nih.gov/ij/plugins/azimuthal-average.html>) was used to integrate the intensity values surrounding the outside of the motor domain with 1° bin width. This generated a plot with two peaks corresponding to the intensity for the stalk and GFP. Fitting using the sum of two Gaussian functions (Igor Pro 6.3, <http://www.wavemetrics.com/products/products.htm>) was used to measure the angle between them. All experiments were done in triplicate. The angle distribution, using a 10° bin width, was visualized by either histogram or rose plot. The number of particles used were as follows: WT + ADP: 3151, 9914, 7534; WT + ADP.Vi: 8710, 5793, 5284; ΔAAA2L H2 + PS-I 8664, 3408, 9220; ΔAAA4L PS-I 9835, 10799, 9955; R1413A + E2028A: 7994, 2445, 11581; ΔAAA3L H2-S3: 6535, 12185, 6399.

Microtubule gliding assays. A microtubule gliding assay was adapted from previously published work⁴². Briefly, anti-GFP antibody (Roche) was non-specifically bound to the glass surface of a flow chamber. The free surface was blocked with assay buffer (30 mM HEPES pH 7.2, 2 mM MgOAc, 1 mM EGTA, 10% (v/v) glycerol, 1 mg ml^{–1} casein and 20 μM paclitaxel). GFP-tagged dynein-2 was then applied and after 30 s incubation washed with assay buffer. Finally, motility buffer that contained microtubules, oxygen scavenging system and 1 mM ATP as an energy source was applied and gliding was observed by total internal reflection fluorescence microscopy. All data analysis used ImageJ⁴³.

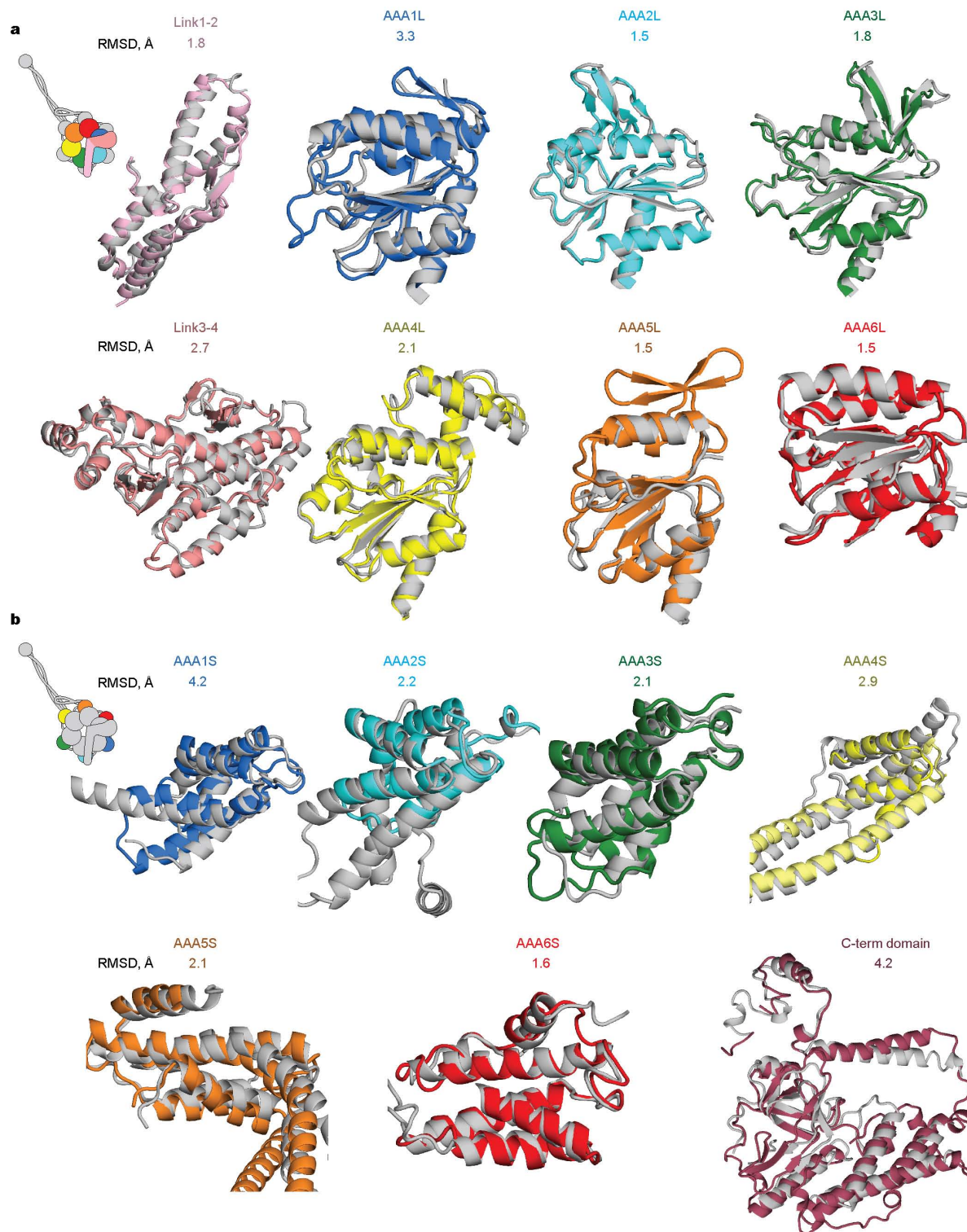
29. Battye, T. G., Kontogiannis, L., Johnson, O., Powell, H. R. & Leslie, A. G. iMOSFLM: a new graphical interface for diffraction-image processing with MOSFLM. *Acta Crystallogr. D* **67**, 271–281 (2011).

30. Evans, P. Scaling and assessment of data quality. *Acta Crystallogr. D* **62**, 72–82 (2006).
31. Strong, M. *et al.* Toward the structural genomics of complexes: crystal structure of a PE/PPE protein complex from *Mycobacterium tuberculosis*. *Proc. Natl Acad. Sci. USA* **103**, 8060–8065 (2006).
32. Vonrhein, C., Blanc, E., Roversi, P. & Bricogne, G. Automated structure solution with autoSHARP. *Methods Mol. Biol.* **364**, 215–230 (2007).
33. Kelley, L. A. & Sternberg, M. J. Protein structure prediction on the Web: a case study using the Phyre server. *Nature Protoc.* **4**, 363–371 (2009).
34. Murshudov, G. N. *et al.* REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr. D* **67**, 355–367 (2011).
35. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
36. Brunger, A. T. *et al.* Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
37. Wallace, A. C., Laskowski, R. A. & Thornton, J. M. LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. *Protein Eng.* **8**, 127–134 (1995).
38. Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M. & Barton, G. J. Jalview version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191 (2009).
39. Huecas, S. & Andreu, J. M. Energetics of the cooperative assembly of cell division protein FtsZ and the nucleotide hydrolysis switch. *J. Biol. Chem.* **278**, 46146–46154 (2003).
40. Tang, G. *et al.* EMAN2: an extensible image processing suite for electron microscopy. *J. Struct. Biol.* **157**, 38–46 (2007).
41. Scheres, S. H. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180**, 519–530 (2012).
42. Reck-Peterson, S. L. *et al.* Single-molecule analysis of dynein processivity and stepping behavior. *Cell* **126**, 335–348 (2006).
43. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nature Methods* **9**, 671–675 (2012).



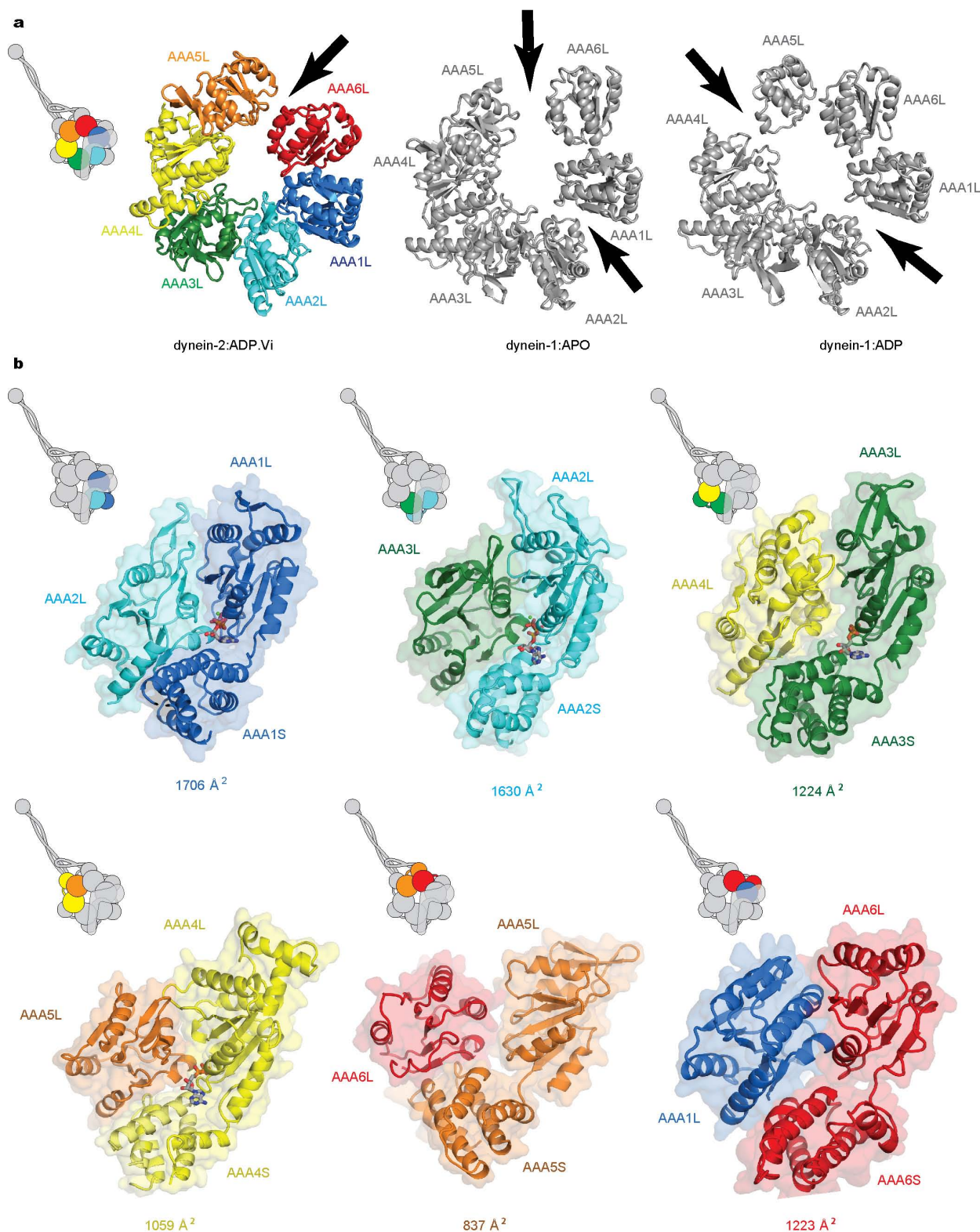
Extended Data Figure 1 | Examples of the electron density quality in dynein-2:ADP.Vi. $2F_o - F_c$ electron density in different parts of dynein-2:ADP.Vi. Amino-acid side-chains are clearly resolved in the linker (a),

AAA1 (b), AAA4 (c) and AAA6 (d). Only the main-chain could be traced in the stalk (e) and the buttress (f). The electron density in a–d was map-sharpened. The contour level is 1σ , except for e which was contoured at 0.75σ .



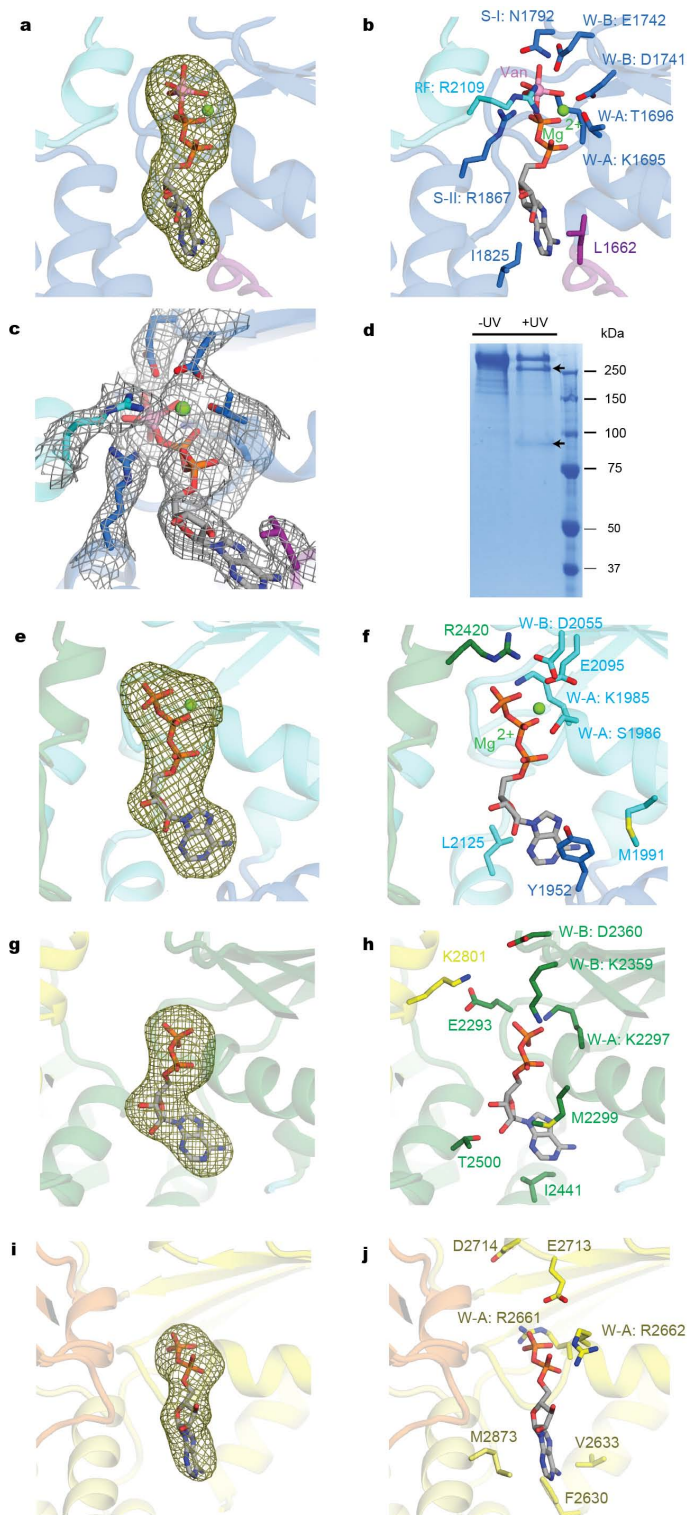
Extended Data Figure 2 | Structural similarity between individual subdomains of dynein-1 and dynein-2. Alignment of individual subdomains from dynein-2:ADP.Vi and dynein-1:ADP (PDB accession number 3VKG). **a**, Alignment of AAA+ large (AAA1L-AAA6L) subdomains and the linker subdomains (Link1-2, Link3-4). **b**, Alignment of individual AAA+ small subdomains (AAA1S-AAA6S) and the C-terminal domain. Dynein-2 subdomains are coloured according to the scheme used in the main text, and shown in the inset cartoons. Dynein-1 subdomains are shown in grey.

Calculated root mean squared deviation (r.m.s.d.) values are shown above each alignment and demonstrate that the subdomains of dynein-2 are structurally highly similar to dynein-1. The AAA+ ring subdomains with the largest r.m.s.d. differences are AAA1L and AAA1S. These subdomains are the most strongly conserved part of the dynein structure and the differences are probably due to the ADP.Vi binding. The distortion of AAA1L, by its interaction with the AAA2L inserts, is described in the main text.

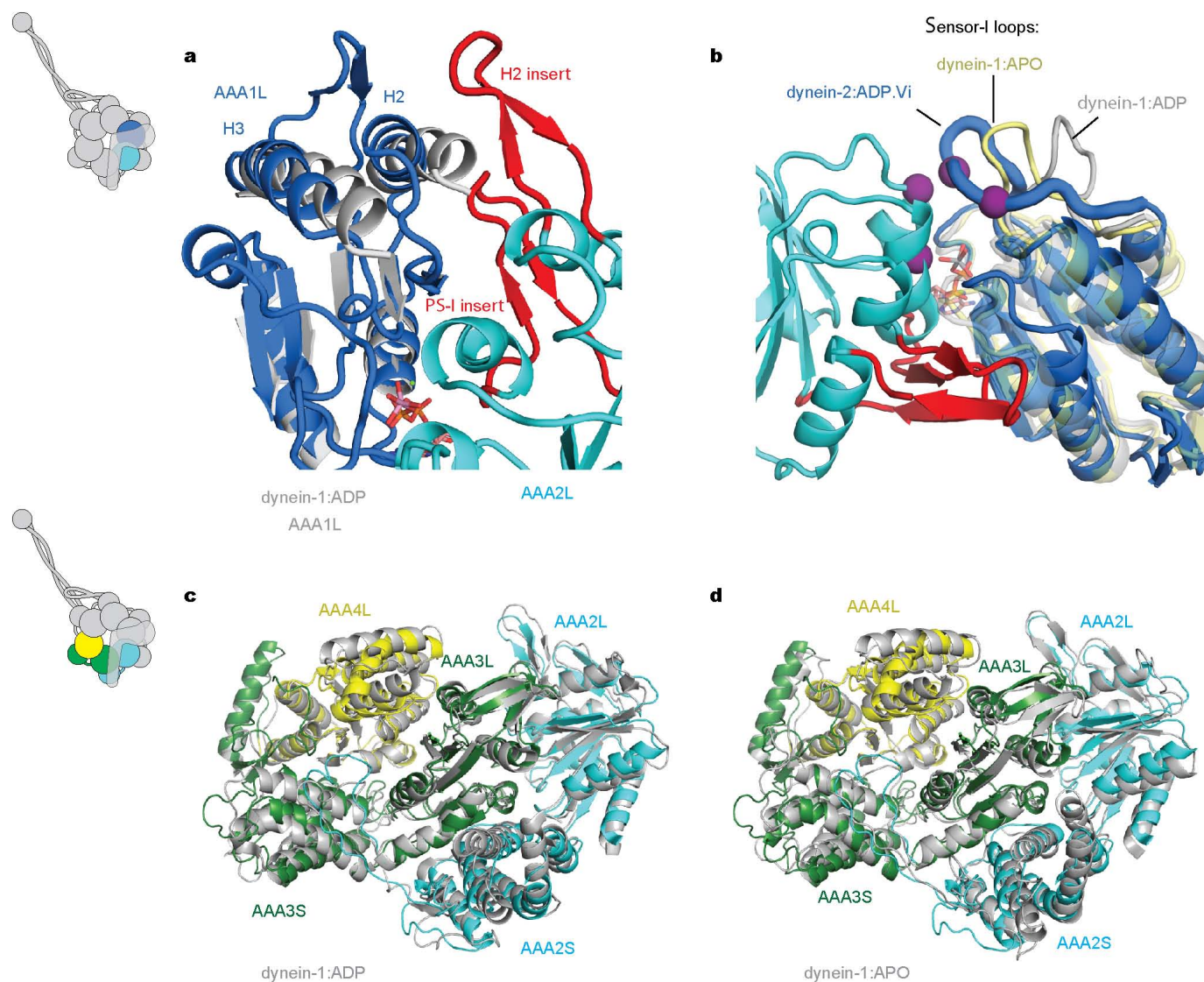


Extended Data Figure 3 | Closed interfaces between AAA+ domains of the AAA+ ring in dynein-2:ADP.Vi. **a**, Gaps in the AAA+ rings of different dynein motor domain crystal structures. In dynein-1:APO (PDB accession number 4AKI) and dynein-1:ADP (PDB accession number 3VKG) there are gaps between AAA1L/AAA2L and AAA5L/AAA6L or AAA4L/AAA5L. In dynein-2:ADP.Vi a smaller gap exists between AAA5L/AAA6L. Gaps are

indicated by black arrows. **b**, Calculated buried surface areas indicate that the interfaces between AAA1/AAA2, AAA2/AAA3, AAA3/AAA4, AAA4/AAA5 and AAA6/AAA1 are tightly closed in dynein-2:ADP.Vi (buried surface areas 1,059–1,706 Å²). The AAA5/AAA6 interface is more open (buried surface area 837 Å²). Nucleotides are shown in stick representation. AAAL, AAA+ large subdomain; AAAS, AAA+ small subdomain.



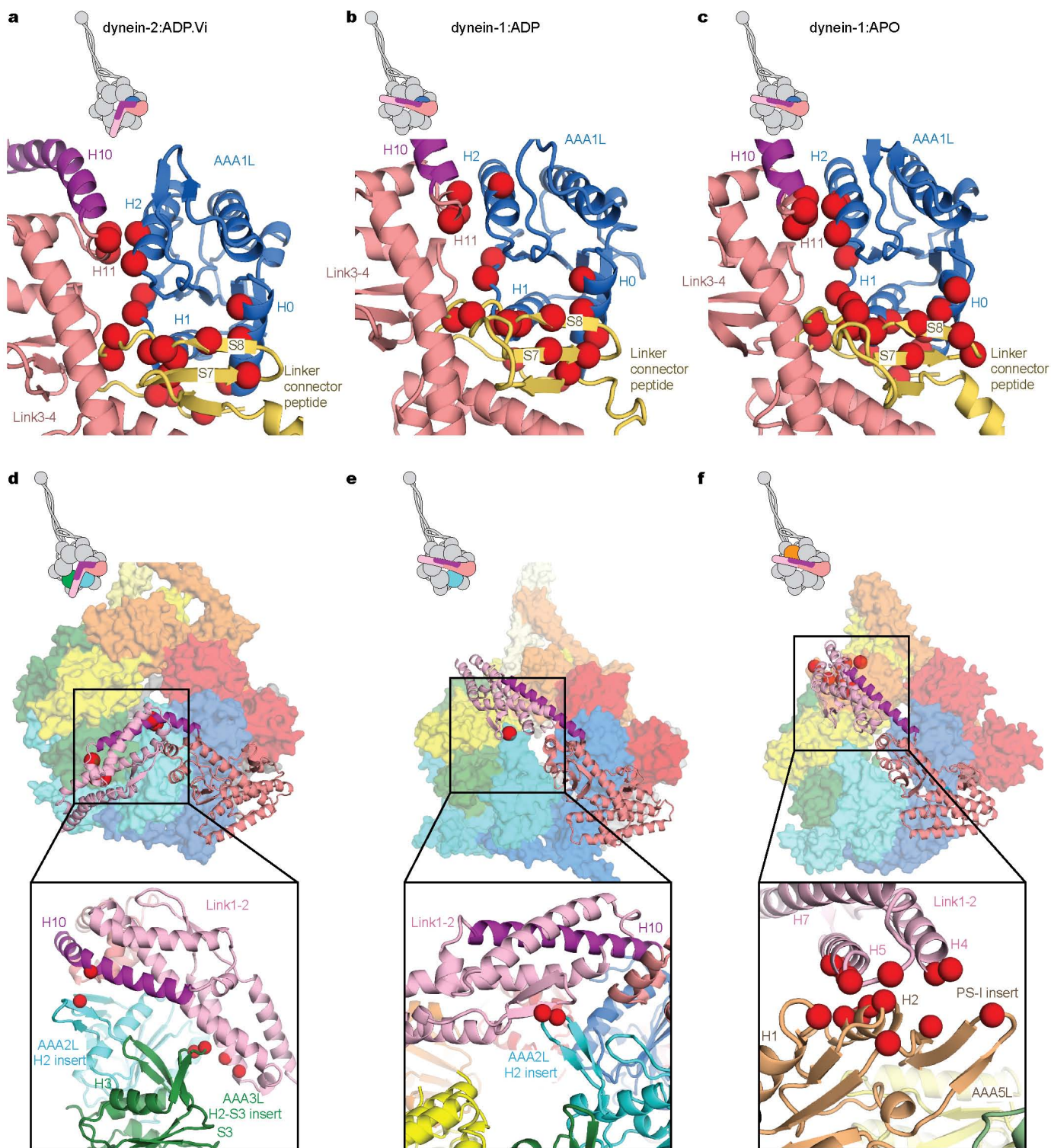
Extended Data Figure 4 | The four nucleotide binding sites of dynein-2:ADP.Vi. **a–c**, The AAA1 site contains electron density consistent with an Mg.ADP.Vi molecule. All catalytic amino-acid residues have the correct conformation to support catalysis. **d**, Photo cleavage¹¹ of washed dynein-2 crystals upon exposure to ultraviolet light (+UV) produces two bands of 300 and 90 kDa (arrowheads). This suggests crystals contain an ADP.Vi group in AAA1. **e, f**, The AAA2 site contains density consistent with a Mg.ATP molecule. **g, h**, The AAA3 and **i, j**, AAA4 sites contain electron density that is best modelled as ADP. In contrast to AAA1, AAA2–AAA4 have lost the catalytic residues necessary for ATP hydrolysis (the Walker B glutamate, the arginine finger, sensor-I and sensor-II motifs). The $F_o - F_c$ electron density (**a, e, g, i**) is contoured at 3σ . The $2F_o - F_c$ electron density (**c**) is contoured at 1σ . W-A, Walker A motif; W-B, Walker B motif; S-I, sensor-I; S-II, sensor-II; RF, arginine finger. Magnesium ions (Mg^{2+}) are shown as green spheres. The vanadium ion of the vanadate molecule (Van) is shown as a pink sphere.



Extended Data Figure 5 | Changes in conformation within dynein

AAA+ ring. **a**, Superimposition of the AAA1L domains of dynein-2:ADP.Vi (blue) and dynein-1:ADP (grey) shows that helices H2 and H3 of AAA1L are displaced when the H2- β hairpin insert of AAA2L (red) comes into contact with H2 of AAA1L. **b**, An alignment of the AAA1L domains of dynein-2:ADP.Vi (blue), dynein-1:APO (PDB accession number 4AKI) (pale yellow)

and dynein-1:ADP (PDB accession number 3VKG) (grey) shows that the loop containing the sensor-I residue is highly variable between the structures. In the presence of ADP.Vi the loop makes contacts (purple spheres) with AAA2L. **c**, **d**, Superimposition of AAA2-AAA4 domains between dynein-2:ADP.Vi and dynein-1:ADP (**c**) or dynein-1:APO (**d**) shows that AAA2-AAA4 move as a rigid body.



Extended Data Figure 6 | Linker interaction with the AAA+ ring in dynein-2:ADP.Vi, dynein-1:ADP and dynein-1:APO. **a–c**, Link3–4 interacts with AAA1L in all structures similar. Mainly hydrophobic contacts exist between the linker H11 helix and the H2 helix as well as the S2 β -sheet of AAA1L. In addition the long peptide that connects the linker with AAA1

(yellow) mediates contacts between Link3–4 and AAA1L. **d**, Link1–2 is stabilized by contacts with AAA2 and AAA3 in dynein-2:ADP.Vi Link1–2, **e** by contacts with the AAA2 H2 insert in dynein-1:ADP and **f** by contacts with AAA5 in dynein-1:APO. Red spheres represent contacts.

Link1-2				Link1-2				Link3-4																												
H5				H7				H10				H11	H12																							
<i>Hs_Cyt-2</i>	1308	LLQSLKDS	1315	1349	WVYLEPIFG	1357	1413	RS	INEE	FLEE	KRS	SA	FP	RF	YFI	GDD	DL	LEI	1440																	
<i>Cr_Cyt-2</i>	1371	LVASLKQS	1378	1412	WVYLEPIFG	1420	1476	R	AL	AD	FLEE	KRS	Q	FP	RF	YFI	FLG	DD	DL	LEI	1503															
<i>Tt_Cyt-2</i>	1300	LLASMKE	1307	1341	WVYLEPIFG	1349	1405	K	AL	N	D	FLEE	KRS	K	FP	RF	YFI	FLG	DD	DL	LEI	1432														
<i>Hs_Cyt-1</i>	1503	SVSAMKLS	1510	1544	WVYLEGIFT	1552	1613	K	AL	GEY	L	ER	S	FP	RF	YFI	FV	G	D	E	D	LEI	1640													
<i>Dd_Cyt-1</i>	1571	SISAMKMS	1578	1612	WVYLEGIFS	1620	1681	K	AL	GEY	L	ER	Q	R	S	A	F	A	R	F	Y	F	V	G	D	E	D	LEI	1708							
<i>Sc_Cyt-1</i>	1411	ELVSMKAS	1418	1452	WLDLYGILG	1460	1521	S	S	L	S	T	F	L	E	R	Q	R	R	Q	F	P	R	F	Y	F	L	G	N	D	D	L	L	K	I	1548
<i>Dm_Cyt-1</i>	1493	SVAAMKLS	1500	1534	WVYLEGIFS	1542	1603	K	AL	GEY	L	ER	R	T	S	F	P	R	F	Y	F	V	G	D	E	D	LEI	1630								
<i>En_Cyt-1</i>	1538	SLQAMRHS	1545	1579	WVYLEGVFT	1587	1648	K	AL	GEY	L	ER	R	V	S	F	P	R	F	Y	F	V	G	D	E	D	LEI	1675								
<i>Ca_Cyt-1</i>	1433	ALTSMKNS	1440	1474	WLYLEGVFG	1482	1544	K	S	L	T	D	Y	L	E	K	Q	R	E	L	F	P	R	F	Y	F	I	G	N	E	D	L	L	E	L	1571
<i>Hs_IDA4_1</i>	1203	MTQNSMFS	1210	1244	WLYLEPIFS	1252	1312	K	G	L	S	E	Y	L	E	T	K	R	S	A	F	P	R	F	Y	F	L	S	D	D	E	L	L	E	I	1339
<i>Hs_IDA3_3</i>	1049	KTQTMCGS	1056	1090	WLYLEPIFS	1098	1158	K	G	L	N	D	Y	L	E	K	K	R	L	F	F	P	R	E	F	F	L	S	N	D	E	L	L	E	I	1185
<i>Hs_ODAg_5</i>	1591	LLGSLLSN	1598	1632	WIYLEAVFV	1640	1701	K	S	L	T	G	Y	L	E	K	K	R	L	C	F	P	R	E	F	F	V	S	D	P	A	L	L	E	I	1728
<i>Hs_IDA5_6</i>	1067	NVATLASS	1074	1108	WLYLESIFN	1116	1176	K	C	L	E	A	Y	L	S	K	R	V	I	F	P	R	F	Y	F	L	S	N	D	E	L	L	E	I	1203	
<i>Hs_IDA3_7</i>	946	KTQTMGRS	953	987	WLYLEPIFS	995	1055	K	G	L	N	E	Y	L	E	K	K	R	L	F	F	P	R	E	F	F	L	S	N	D	E	L	L	E	I	1082
<i>Hs_ODAg_8</i>	1457	VLGSLLSN	1464	1498	WVYLEAVFV	1506	1567	K	S	L	T	G	Y	L	E	K	K	R	L	L	F	P	R	E	F	F	V	S	D	P	V	L	L	E	I	1594
<i>Hs_ODAab_9</i>	1478	QLQNLVMS	1485	1519	WTHLESIFT	1527	1588	K	A	A	E	Y	L	D	T	K	R	L	A	F	P	R	F	Y	F	L	S	S	D	L	L	D	I	1615		
<i>Hs_ODAab_17</i>	1451	QLQNLMS	1458	1492	WSHLESIFI	1500	1561	K	A	A	E	Y	L	E	T	K	R	L	A	F	P	R	F	Y	F	V	S	S	A	D	L	L	D	I	1688	

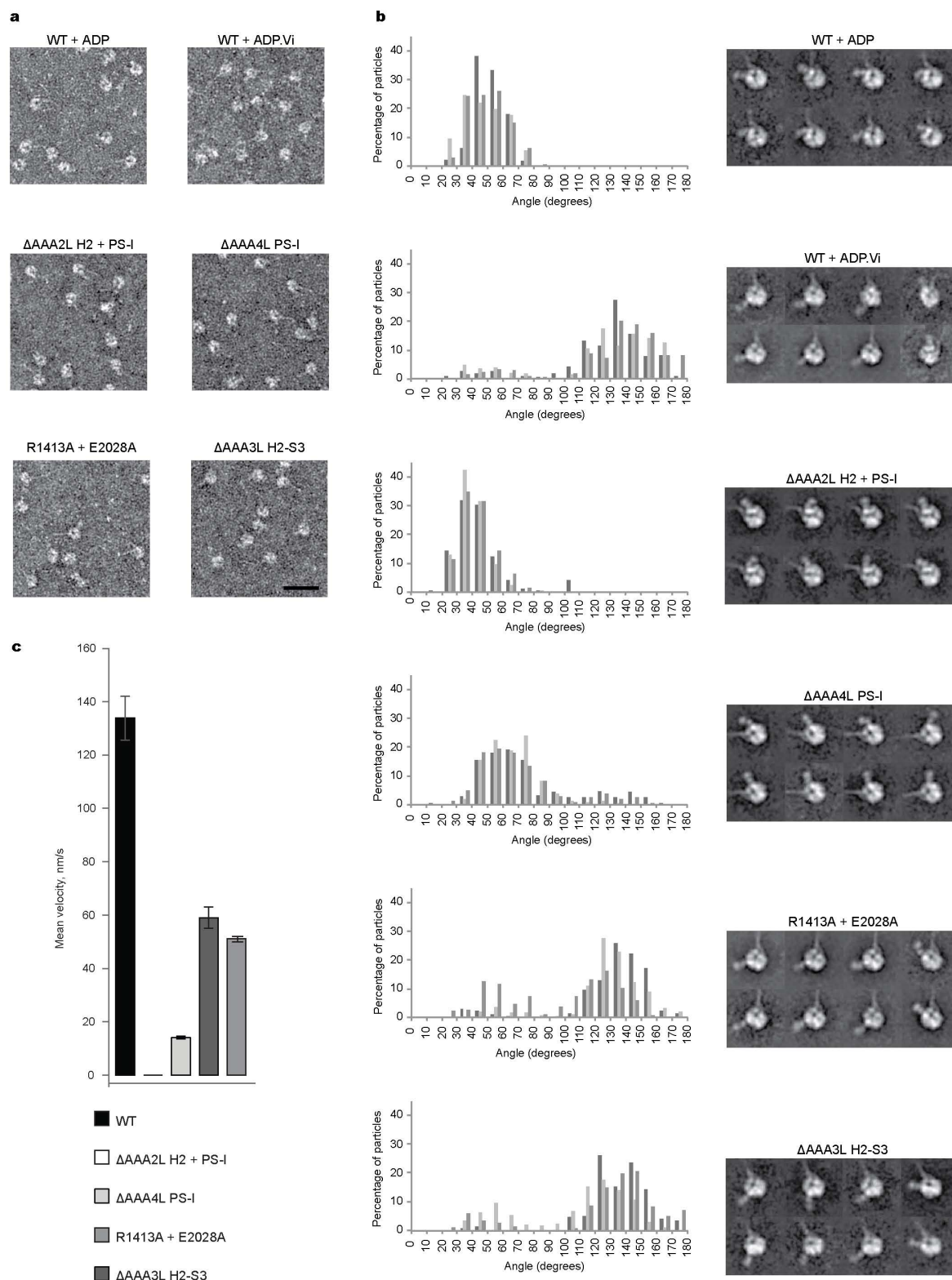
AAA2L H2 insert

AAA3L H2-S3 insert

<i>Hs_Cyt-2</i>	2018	LLGHIDMTREWS	DGVL	2034	2335	CMV-ISTNTGRVYRPK-DCER	2353
<i>Cr_Cyt-2</i>	2073	LLGSMNMDTREWS	DGVL	2089	2378	CGKPVTTTSGKALRPP-DNTR	2397
<i>Tt_Cyt-2</i>	1999	LLGLMNNDTREFTE	GV	2015	2317	CMK-GTFSQGRILKPK-DASR	2335
<i>Hs_Cyt-1</i>	2264	LYGTLDPNTRWTD	DGLF	2280	2639	CEYRRTPNGV-VLAPVQLGKW	2658
<i>Dd_Cyt-1</i>	2316	LFGSLDLTTREWTD	DGLF	2332	2718	CEYKRTPSGETVLRPTQLGKW	2738
<i>Sc_Cyt-1</i>	2114	LYGSMKATLEWRD	DGLF	2130	2462	TNYVTTSKGL-TLLPKSDIKN	2481
<i>Dm_Cyt-1</i>	2250	LYGVLDPNTRWTD	DGLF	2266	2624	CEYRTPNGV-VLSPVQIGKW	2643
<i>En_Cyt-1</i>	2269	LYGSLDSTTREWTD	DGLF	2285	2637	CEYKKTLSGV-VMSPNQIGRW	2656
<i>Ca_Cyt-1</i>	2152	IYGKLDLVTRDWT	DGLF	2168	2511	CEYRKTNRGI-QLAPRINGKW	2530
<i>Hs_IDA4_1</i>	1911	LYGEFDLLTHEWTD	IGIF	1927	2269	IDSKLDKRRKG VFGPP-LGRN	2288
<i>Hs_IDA3_3</i>	1756	LYGCFDQVSHWMD	GV	1772	2116	IMSKLDRRRKGLFGPP-IGKK	2135
<i>Hs_ODAq_5</i>	2301	MFGRLDVATINDWT	IGIF	2317	2632	IESYVDKRMGT TYGPP-AGKK	2651
<i>Hs_IDA5_6</i>	1801	LYGEVNNLTLEWKD	GGLM	1817	2139	IESKLERKRNILGAP-GNKR	2158
<i>Hs_IDA3_7</i>	1655	LYGQFDSVSHWSD	GV	1671	2019	VMSKLDKRRKG VFGPP-LGKR	2038
<i>Hs_ODAg_8</i>	2167	MFGRLDTATNDWT	IGIF	2183	2498	IESYVDKRIGSTY GPP-GGRK	2517
<i>Hs_ODAab_9</i>	2190	LFGIINPATGEWKD	GLF	2206	2520	LEKPLEKKAGRNYGPP-GNKK	2539
<i>Hs_ODAab_17</i>	2153	LFGIINPVTREWKD	GLF	2169	2483	LEKPLEKKSGRNYGPP-GTKK	2502

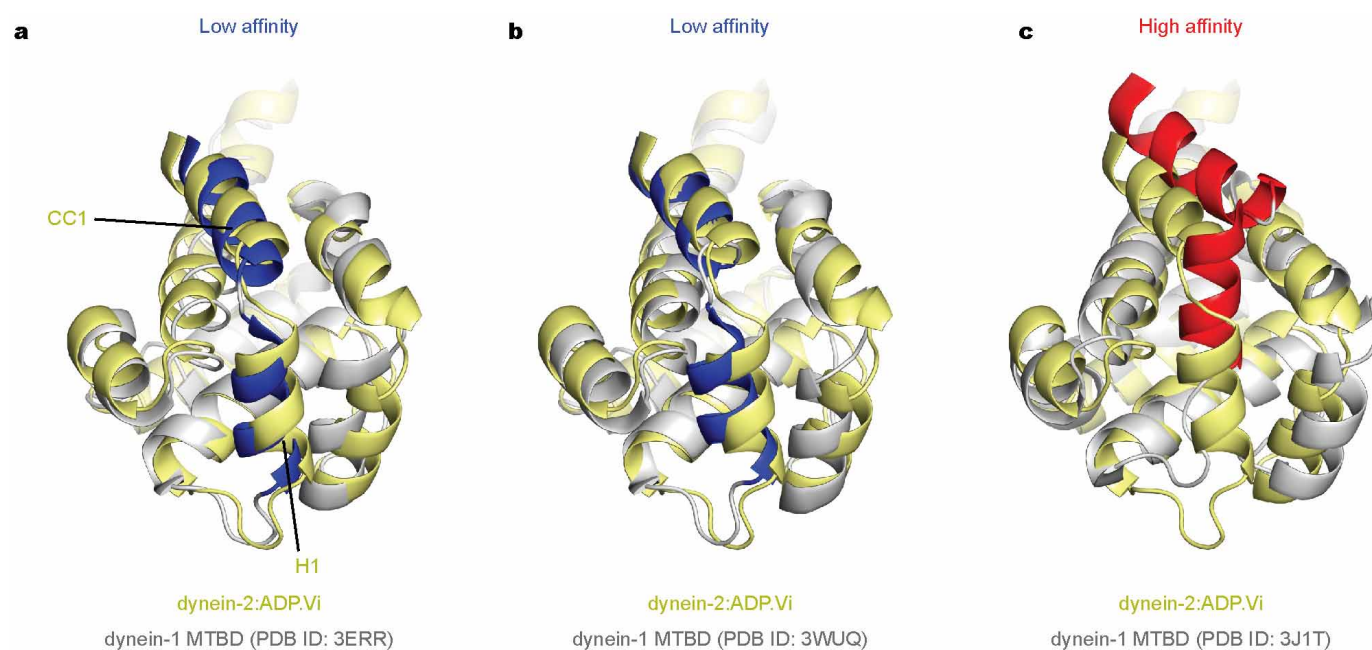
Extended Data Figure 7 | Conservation of contact sites between linker and dynein ring. Multiple alignment of cytoplasmic dynein-1 (Cyt-1), dynein-2 (Cyt-2), axonemal inner arm dyneins (IDA) and outer arm γ (ODAg) and $\alpha\beta$ (ODAab) dyneins. Dyneins are from human (Hs), *Chlamydomonas reinhardtii* (Cr), *Tetrahymena thermophila* (Tt), *D. discoideum* (Dd), *S. cerevisiae* (Sc), *Drosophila melanogaster* (Dm), *Emmericella nidulans* (En) and

Candida albicans (Ca). Residues are shaded by conservation, with dark blue being the most conserved. Red asterisks mark hydrophobic contacts that stabilize the bent linker conformation, black asterisks mark the contact site between AAA2L H2 insert (E2028) and the linker (R1413) and green asterisks mark poorly conserved contacts between the linker and the AAA3 H2-S3 insert.



Extended Data Figure 8 | Characterization of dynein-2 mutants by negative electron microscopy and microtubule gliding assays. **a**, Representative micrographs showing the quality of the raw electron microscopy data. Scale bar, 20 nm. **b**, Left, histograms showing distribution of angles between the linker and the stalk in three replicate negative-stain electron microscopy experiments (10° bin width); right, representative subclasses used for angle measurement. **c**, Mean velocities of dynein-2 mutants in microtubule gliding assays.

GFP-dynein-2_{D1091-Q4307} (wild type: WT) glides microtubules at $134 \pm 8 \text{ nm s}^{-1}$ ($N = 99$). The microtubule gliding velocities for the other constructs are Δ AAA3L H2-S3, $59 \pm 4 \text{ nm s}^{-1}$ ($N = 79$); K1413A + E2028A, $49 \pm 2 \text{ nm s}^{-1}$ ($N = 31$); and Δ AAA4L PS-I, $14 \pm 1 \text{ nm s}^{-1}$ ($N = 121$). Microtubule gliding was not observed in case of Δ AAA2L H2 + PS-I. Error bars, s.e.m.



Extended Data Figure 9 | The MTBD in dynein-2 ADP.Vi is in the low microtubule affinity conformation. **a, b,** Alignment of dynein-2 ADP.Vi MTBD (pale yellow) with dynein-1 MTBDs (grey) in the low microtubule affinity conformation (PDB accession numbers 3ERR and 3WUQ respectively), and **c** with a dynein-1 MTBD in the high microtubule affinity conformation (PDB accession number 3J1T). The stalk CC1 and the MTBD H1

undergo conformational changes depending on the microtubule affinity of the MTBD. In dynein-2:ADP.Vi the arrangement of these structural elements suggests the MTBD is in the low microtubule affinity conformation. Stalk CC1 and MTBD H1 are coloured blue in low-affinity structures and red in high-affinity structures.

Extended Data Table 1 | Data collection, phasing and refinement statistics

Dataset	Native-1	Native-2	Na ₃ [PW ₁₂ O ₄₀]	
Space group	C222 ₁	C222 ₁	C222 ₁	
Cell dimensions				
<i>a</i> , <i>b</i> , <i>c</i> (Å)	136.0, 487.2, 276.5	136.2, 487.7, 276.9	135.7, 481.9, 276.5	
			<i>Peak</i>	<i>Inflection</i>
Wavelength (Å)	0.97949	0.97949	1.21416	1.21476
Resolution (Å)	56.5-3.40	56.1-6.0	65.9-6.0	69.5-6.0
<i>R</i> _{sym} or <i>R</i> _{merge}	10.1 (69.2)*	6.3 (17.7)	24.0 (111.6)	36.1 (170.6)
<i>I</i> / <i>σ</i> <i>I</i>	7.6 (1.1)	207.2 (28.4)	8.0 (2.4)	5.9 (1.7)
Completeness (%)	62.2 (1.9) [§]	94.9 (99.6)	99.8 (99.9)	99.8 (100.0)
Redundancy	4.1	3.6	11.7	10.0
Refinement				
Resolution (Å)	56.6-3.41			
No. reflections	74060			
<i>R</i> _{work} / <i>R</i> _{free}	23.7/28.5			
No. atoms	22816			
Protein	22697			
Ligand/ion	119			
Water	-			
B-factors				
Protein	122.0			
Ligand/ion	69.8			
Water	-			
R.m.s deviations				
Bond lengths (Å)	0.012			
Bond angles (°)	1.55			

*Highest resolution shell is shown in parenthesis.

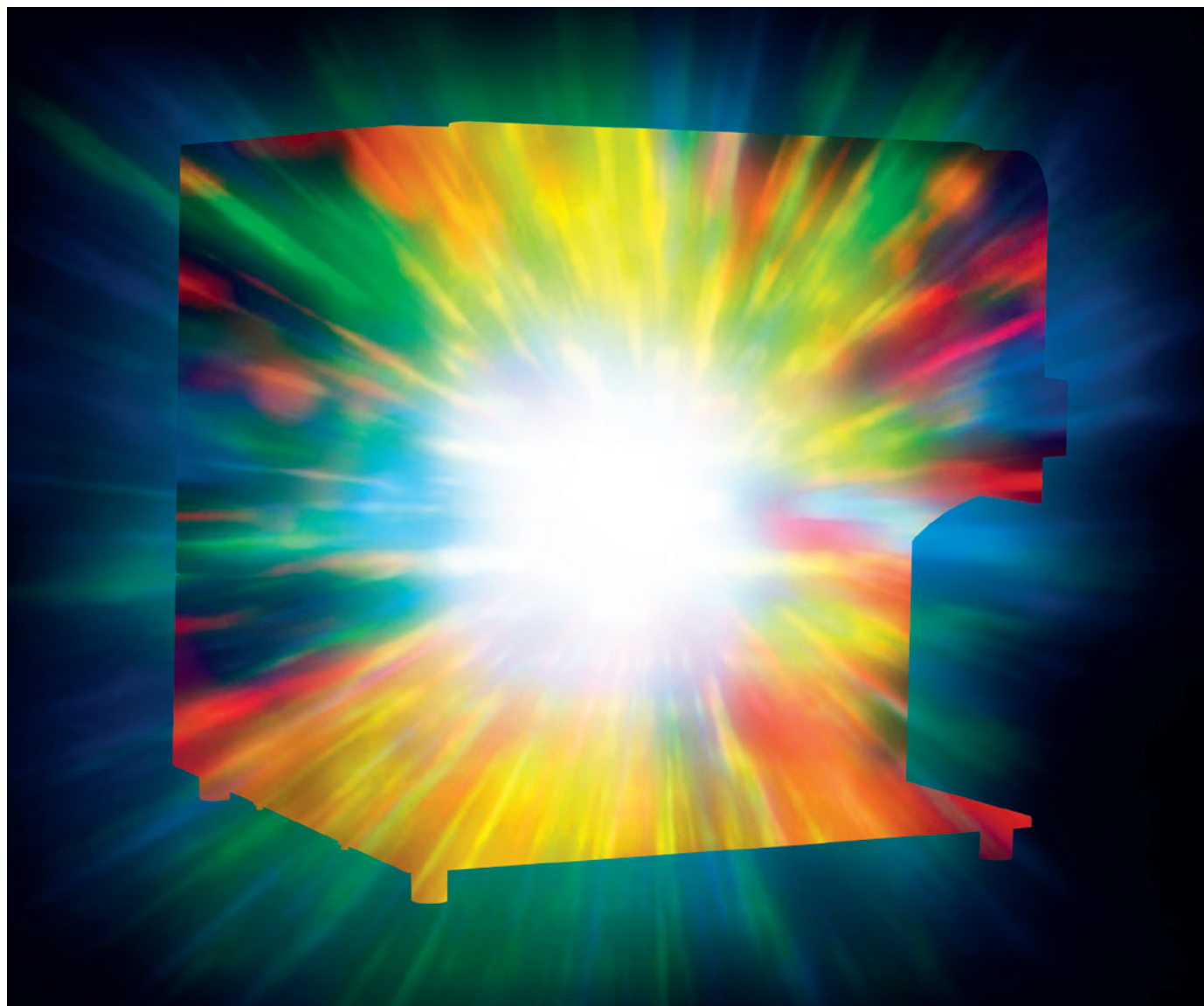
[§]The low completeness to 3.4 Å is due to the anisotropic diffraction limits (*a*= 4.2 Å, *b*= 4.4 Å, *c*= 3.4 Å). When the data is scaled to a resolution limit of 4.2 Å the overall completeness is 97.9% and the completeness in the highest resolution shell is 78.4%.

TECHNOLOGY FEATURE

MEASURE FOR MEASURE

Cutting-edge tools that can identify the characteristics of cells are helping researchers to develop more-effective vaccines.

BD BIOSCIENCES



BY JIM KLING

Vaccines are a triumph of science over infection. They have defeated smallpox, which the World Health Organization declared eradicated in 1980, and dramatically lowered the toll of many other infectious diseases.

But not all. The search for vaccines against conditions such as HIV/AIDS and malaria

has been hindered by researchers' incomplete understanding of the human immune system, says Mario Roederer, an immunologist at the US National Institutes of Health's Vaccine Research Center in Bethesda, Maryland.

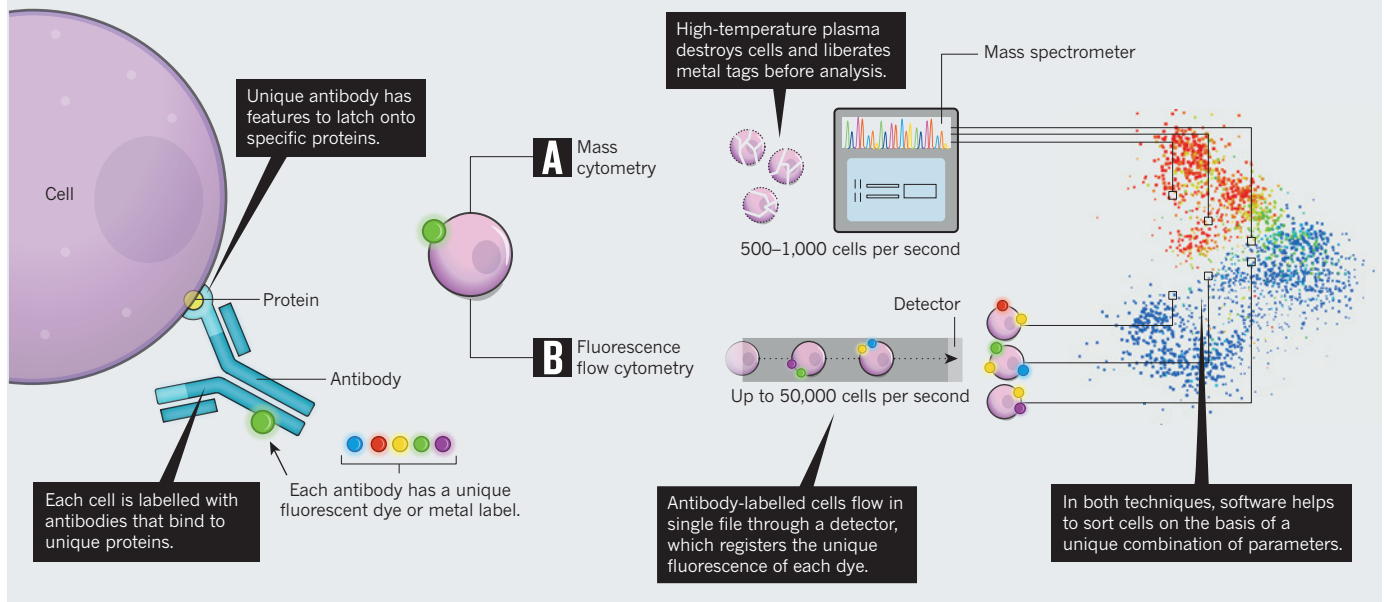
In general, says Roederer, researchers know that a successful vaccine jump-starts antibody production and other defences. But they do not know which of the immune system's thousand or more functional cell types direct

the response against individual pathogens. If researchers could identify those cell types, he says, then they could devise vaccines that maximize production of these cells.

One of the best ways to identify those cells is flow cytometry, a technique that analyses and sorts cells according to their distinguishing characteristics — usually proteins on the outer surface — and reveals much about a cell's function and position in the immune ►

GO WITH THE FLOW

Fluorescence flow cytometers and mass cytometers use antibodies to interrogate up to 50,000 cells per second. Cells are passed through a detector or analysed with a mass spectrometer to sort them by shared characteristics.



► system. But current-generation flow cytometry is detailed enough to place cells only in broad categories — like identifying something simply as a fish instead of as a great white shark. Researchers know that there is a specific predator that they want, says Roederer. “We’re just trying to figure out how to find it.”

The ability to identify specific cells may also improve researchers’ understanding of diseases such as multiple sclerosis, in which the immune system attacks the host’s own tissues, and metastatic cancer, in which rogue cells from the original tumour migrate into other tissues.

Such possibilities have motivated researchers to develop two new approaches, each of which promises to double or even triple the

limit of conventional techniques by 2016 (see ‘Colour bursts’).

One approach is a variant of standard flow cytometry that uses a new type of very intense fluorescent dye and can identify 27 proteins. The other, known as mass cytometry, can record 50 parameters such as cell-surface proteins or parts of proteins, says immunologist Garry Nolan at Stanford University in California. It labels the proteins with metal atoms, then records the weight of the atoms within each cell with a mass spectrometer to provide a signature for each cell type. The labels can also help in microscope imaging of tissues. For instance, they show where cell-surface proteins are located in a slice of excised tumour. That adds layers of information about the cell types present in the tumour and hints at their function within the cancerous growth.

WIDEN THE NET

Like the immune system, flow cytometry uses antibodies to seek out proteins (see ‘Go with the flow’). First, researchers create an antibody for each protein they want to study, then they label the antibody with a dye molecule that can absorb light and be made to fluoresce in a specific colour.

The sample to be studied is then bathed in the labelled antibodies, which stick to the cells that bear the matching proteins. The antibody-decorated cells are then directed one at a time through a narrow channel. As they pass through, a light pulse triggers the dyes to flash, revealing which proteins are present on the cells. The light from each dye fans out across different wavelengths to produce a readout that looks like a broad mountain peak.

At the moment, flow cytometers can process

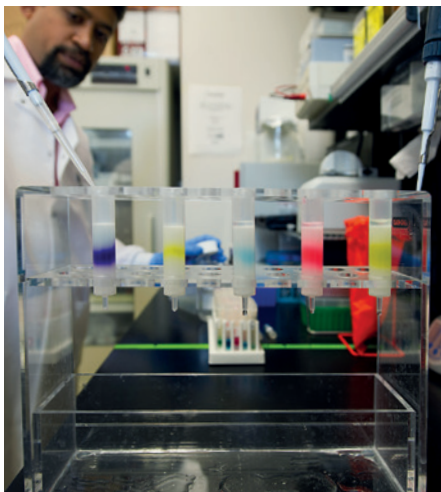
no more than 18 fluorescent dyes at once because when more than this number fluoresce simultaneously, the shoulders of some spectra will overlap with the crests of others, making the crests impossible to discern.

“I’ve always been lucky to be surrounded by the necessary technologies.”

A complicating factor is that the abundance of each protein can vary greatly from cell to cell. A cell that has thousands of a specific protein on its surface will attract many identically labelled antibodies, which combine to produce a bright flash of fluorescence. A rare protein will produce a weaker signal that may get overwhelmed by those from more abundant proteins.

Researchers can compensate by using brighter dyes to label antibodies targeted at the less common proteins, says Roederer. But they often do not know the relative abundances of proteins in advance, he says, so they may have to spend weeks relying on trial and error to work out which dyes to use on which antibodies.

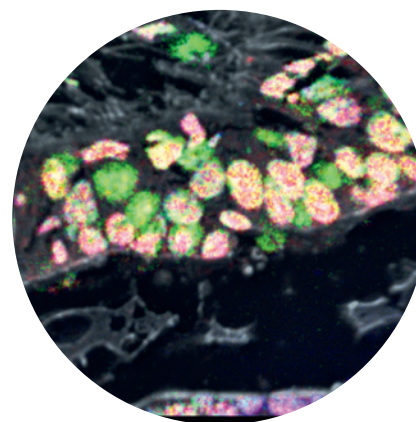
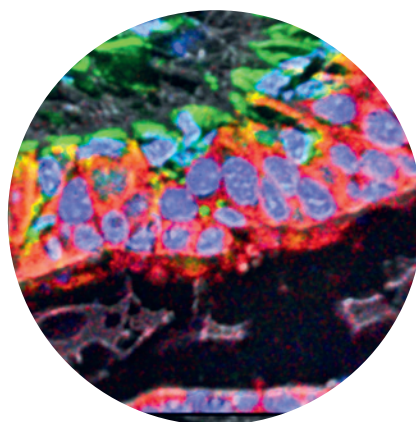
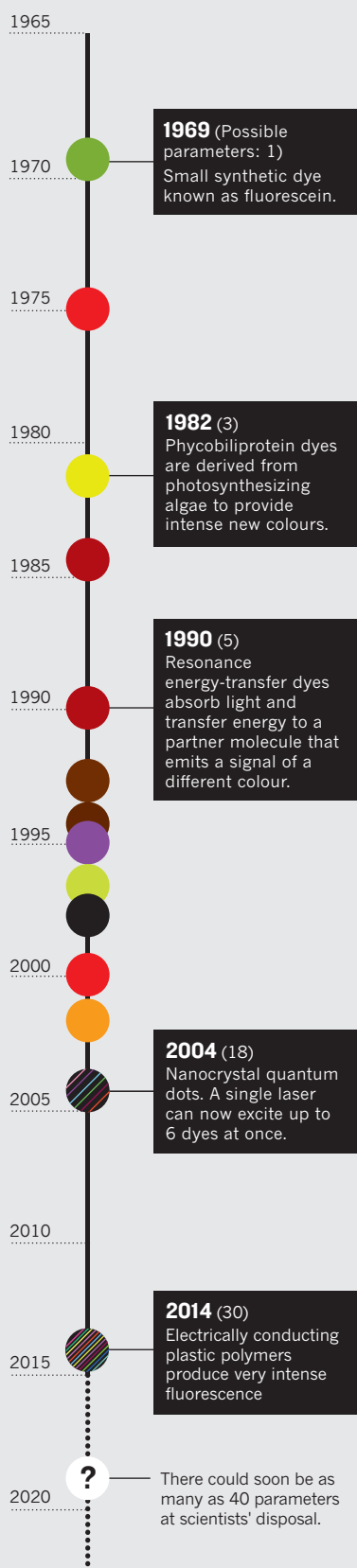
That could well change soon, because researchers have developed a new class of dye¹. Made from electrically conducting plastic polymers, these dyes act as miniature antennas that absorb energy from the light pulse at multiple points that all fluoresce together, producing a more intense signal. When used in flow cytometry, says Roederer, the intense signal overwhelms any overlap from other dyes, even for low-abundance proteins, and means that more dyes can be used simultaneously. “You have so much more flexibility,” says Roederer, whose group has used the dyes to survey 30 proteins commonly found on the



By using antibodies labelled with fluorescent dyes, researchers identify the proteins present on a cell.

COLOUR BURSTS

Every decade, the discovery of new classes of fluorescent dye (coloured dots) roughly doubles the number of simultaneous measurements (parameters) possible per cell.



surfaces of immune cells, although it has yet to publish the results.

Other research teams are using these dyes, which are available from two Californian firms — BD Biosciences in San Jose (where they are marketed as BD Horizon Brilliant dyes), and BioLegend of San Diego (marketed as Brilliant Violet dyes) — but Roederer knows of no one else who has managed to measure more than 18 parameters simultaneously. Measuring more than that requires specialized instruments, software and chemical expertise. “It’s not trivial” to do, he says. “I’ve always been lucky to be surrounded by the necessary technologies.”

Roederer now plans to use the dyes to study vaccine candidates against several diseases, including Ebola, malaria, tuberculosis and HIV/AIDS. An Ebola vaccine currently being tested in human volunteers is a priority. Roederer joined a team led by immunologist Nancy Sullivan, also at the Vaccine Research Center, that narrowed down the cells that protect monkeys from Ebola — not quite to shark level, but to about that of a predatory fish². With more dyes, Roederer expects to find his shark and pinpoint the cell populations that give the monkeys immunity. Then Sullivan and Roederer hope to tune the vaccine doses and schedules to get the human immune system to produce and support those cells.

METALS, NOT COLOURS

As Roederer forges ahead with the new dyes, DVS Sciences of Sunnyvale, California, is taking flow cytometry in a different direction. The company, now called Fluidigm, introduced the first commercial mass cytometer in 2009 and debuted its next-generation machine, the CyTOF 2, in 2013.

Like flow cytometry, mass cytometry involves soaking cells in labelled antibodies then squeezing them into a narrow flow to be screened one by one. But that is where the similarity between the two techniques ends. In place of fluorescent dyes, mass cytometry uses rare earth metals, which are absent from living systems, to label the proteins. And rather than illuminating the cells to reveal the presence of proteins, the mass cytometer

uses high-temperature plasma to split the cells into their component atoms. These atoms, which now include the rare-earth labels, then get fed into a mass spectrometer to measure the mass and abundance of each metal, and thus the identity and abundance of each matched protein. There are therefore no problems with signal overlap as there are in fluorescent dyes.

A group led by Nolan used the technique to look for specific immune-cell populations in patients recovering from hip-replacement surgery — which, like most traumas, prompts a complex response from the immune system to orchestrate healing. Nolan reasoned that he could identify the immune cells that promote faster recovery in some patients. His team collected blood from each patient before their surgery and at various times afterward, and tracked 31 proteins using mass cytometry. In patients who recovered quickly, the researchers found unique types of immune cells known as monocytes³. “We were able to see signatures that were changing and that lined up really well with surgical recovery,” says Sean Bendall, a pathologist at Stanford, and a co-author of the study.

Researchers are now investigating whether it is possible to use these cells to predict which patients are likely to have delays in recovery, and provide interventions that could improve their recovery time, says Bendall.

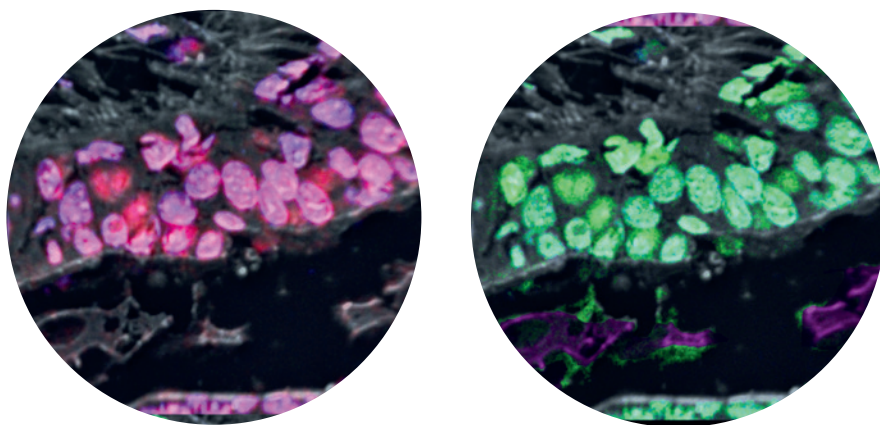
Mass cytometry and fluorescence flow cytometry compete for some of the same research applications, and researchers tend to have a preferred technique.

Roederer points out that mass cytometry destroys cells, whereas fluorescence flow cytometry can preserve them, and can even isolate and sort them at the same time.

Bendall does not share that concern. He enthusiastically backs mass cytometry, but says that he often hears objections about the destruction of cells. “I feel like that’s the ‘gotcha’ question,” he says. Many people will dismiss the technique simply because of that, he explains, but cell destruction “has never been a road-block to anything we wanted to do”. If a mass-cytometry experiment identifies an interesting cell type, he explains, then researchers can

MICHAEL ANGELO, STANFORD UNIV.

SOURCE: MARIO ROEDERER. DESIGN: CLAIRE WELSH/NATURE



Multiplexed ion-beam imaging can reveal different DNA-binding proteins in human breast tissue. Each image highlights four proteins.

different colours to reveal the location of individual proteins within the section⁴.

In a related approach, immunologist Bernd Bodenmiller at the University of Zurich in Switzerland and chemist Detlef Günther at the Swiss Federal Institute of Technology in Zurich and their colleagues have developed a companion instrument. It records the position of the cells and directs an ultraviolet laser at the tissue to methodically strip away labelled cells and send them to a mass cytometer for analysis⁵. The instrument can analyse up to 40 tags, Bodenmiller says. However, there are now more metal tags available, so researchers will probably soon conduct experiments with even more parameters, says Scott Tanner, Fluidigm's chief technical officer.

Images produced by this technique could illuminate how cells respond to local conditions within a tumour, such as low-oxygen environments. "It gives you yet more information about the nature of that sample," says Tanner.

Fluidigm has licensed the laser ablation chamber and given prototypes to several academic research teams. It hopes to begin selling the device this year.

Imaging mass cytometry could be applied beyond cancer, Tanner says. Neurobiologists tell him that they hope to use it to examine how neurons are distributed in the brain or spinal cord. The ability to analyse multiple proteins could help researchers to decipher the functions of neural cells and how they relate to the cell's location within a network of linked neurons.

Roederer and Nolan are constantly pushing the boundaries, but Roederer says that he often encounters scepticism of how useful the improvements are. When he achieved 8 parameters, researchers questioned if that many were really necessary. "When I got to 12, people were saying 'Is that enough? Are you done yet?'" he recalls.

He is not done. He hopes to get to 40 parameters next year, and even more after that. "Every time we reached a new ceiling, we were looking to crack it within a couple of years." ■

Jim Kling is a freelance science writer in Bellingham, Washington.

1. Chattopadhyay, P. K. *et al. Cytometry A* **81A**, 456–466 (2012).
2. Stanley, D. A. *et al. Nature Med.* **20**, 1126–1129 (2014).
3. Gaudillière, B. *et al. Sci. Transl. Med.* **6**, 255ra131 (2014).
4. Angelo, M. *et al. Nature Med.* **20**, 436–442 (2014).
5. Giesen, C. *et al. Nature Meth.* **11**, 417–422 (2014).

always use that information to isolate living counterparts with a traditional flow cytometer.

Fluorescence-based flow cytometry does, however, have the advantage of being faster: it can analyse up to 50,000 cells per second, whereas mass cytometry manages no more than 1,000 per second because the cloud of atoms produced by the plasma takes time to clear before the instrument can accommodate the next cell.

The intense fluorescent dyes are not without their drawbacks. They are so new that little road-testing has been done using high numbers simultaneously — and few papers have been published on them.

That leaves mass-cytometry advocate Nolan a little sceptical of the technique. "I've yet to see them implemented in a way that would have me jump ship and go back to fluorescence," he says.

PICTURES PERFECT

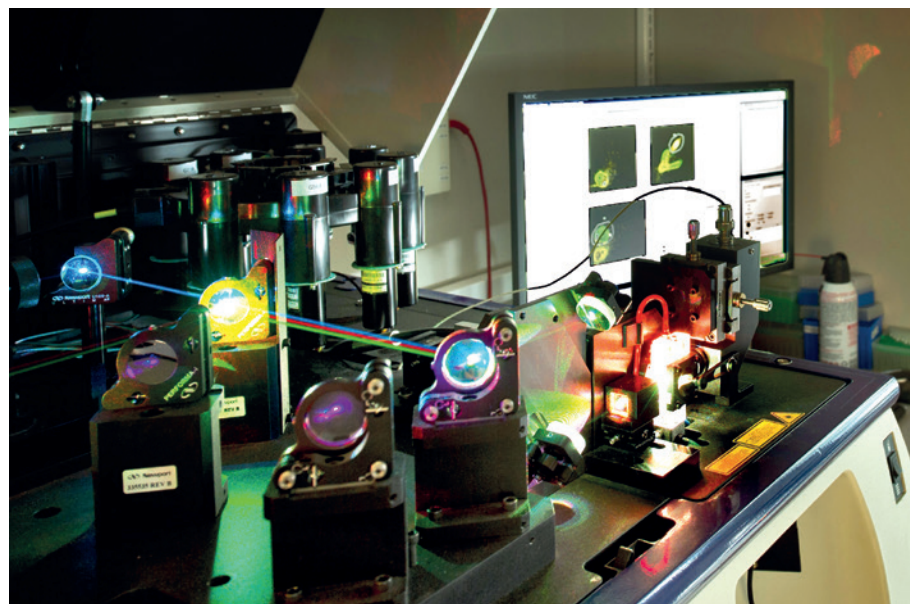
Nolan, Bendall and Roederer all agree on one thing: the most promising application of metal tags is their potential to improve images

of intact slices of tissue — and that cannot be done with either fluorescence flow cytometry or mass cytometry because both techniques require cells to be dispersed into a flowing stream.

One such application, called multiplexed ion-beam imaging, applies metal tags to tissue slices, then showers them with oxygen ions. The oxygen ions react with the metal tags, dislodging them from their accompanying antibodies. A mass spectrometer then measures the metal atoms as they ricochet away from the tissue.

Michael Angelo, a pathologist at Stanford, says he has pushed the technique to 45 parameters per cell. The method also records the cell's position within, say, a tumour. "I think it's extraordinarily cool. That will never be done by traditional imaging or flow-cytometry-based technologies," Roederer says.

Last year, Nolan's group used the technique in tissue samples from people with breast cancer using ten metal-labelled antibodies. The technique produced a high-definition picture of the tissue that could then be displayed in



In fluorescence flow cytometry, lasers stimulate dyes to flash a unique colour and the instrument registers those flashes to characterize each cell.

CAREERS

HOT COLLABORATOR Researcher reaches top rank as middle author **p.447**

REGULATIONS Study on scientific paperwork might cut red tape **p.447**

DATA-DRIVEN DECISIONS An academic uses stats to choose a job **blogs.nature.com/naturejobs**

ANDY BAKER/GETTY



BY CHRIS WOOLSTON

Before each dig, archaeologist Michael Danti makes a mental inventory of everything that might go wrong. His work takes place in Syria and Kurdistan, so the list includes potential crises such as encountering poisonous scorpions, wild dogs, heat stroke, unexploded mines and, lately, armed secret police in search of insurgents. “Getting misidentified by the police is a big danger,” he says. “Archaeologists and ISIS members can dress similarly. It’s important for us to look as Western as possible.”

Danti, who is employed by Boston University in Massachusetts, has heard his share of explosions and gunfire during his fieldwork, but nobody on his team has ever been seriously injured. Over the years, he has learned that staying safe often means watching out for the locals — wherever he may be. Before he focused on the Middle East, he conducted archaeology in the rural US Midwest. “That was just as dangerous,” he says. “You could stumble across somebody’s methamphetamine lab or get shot at by an angry farmer.”

When most scientists think about safety, they focus on equipment, reagents or radiation — not armed militia or shotgun-wielding farmers. Violence of any sort is probably far down their list of concerns.

But no matter their area of study, scientists eventually encounter the human element, a potentially volatile hazard that does not come with operating instructions. From foreign soldiers or campus criminals to unhinged colleagues or team leaders with bad intentions — harm can come from many directions. “It’s like any other job,” says Danti of his work. “The safety end of it is very important.”

Whether at remote field camps or in the lab, scientists and institutions need to pay attention to personal safety, says Michael Dorn, executive director of the campus-safety advocacy group Safe Havens International in Macon, Georgia. “I see a lot of people who are very accomplished but also very reckless,” he says. “You may be a great researcher, but it doesn’t make sense to disregard danger.”

Scientists in every discipline and every setting should apply their analytical skills to the possible hazards around them, Dorn says. “The brain is a very powerful survival mechanism,” he says. As a start, he recommends paying close attention to other people, noting behaviours or activities that seem out of place. “Scientists are very good at recognizing ►

WORKPLACE SAFETY

Risky encounters

Of all of the hazards faced by researchers, the biggest threat may be the human element.

► patterns,” he says. “If something seems odd, they should take that seriously.”

And if junior researchers feel unsafe about some aspect of the job — whether it is late nights at the lab or field work in a dangerous part of the world — they should consult their superiors. “It’s very prudent to ask questions and raise concerns if you have them,” he says. Once questions are asked, the principal investigator (PI) may suggest new safeguards, such as eliminating night-time experiments or that people work in pairs. If nothing else, Dorn says, questions help to keep safety in the conversation.

Whether they are working at a far-flung site or close to home, PIs should have clear and explicit rules for staying safe, and junior researchers need to make sure that they understand the reason for every mandate.

SAFETY FIRST

Danti has a set of safety rules for his digs: nobody works alone, nobody drives at night and junior trainees never drive. He is not concerned about the vehicle; rather, he worries that the person driving could end up being stoned to death if they accidentally hit a villager and are not familiar with local customs and protocols for behaviour. Once lab members understand the rationale, it pretty much eliminates the likelihood of a joy ride. He also has several strategies in place for leaving a country in a hurry, and tries to make sure that whenever it is possible, he has an Internet connection to keep track of the swirling politics and conflicts in any region.

Yet even staying connected can have a downside. Having online access, he points out, means that some team members may be too distracted by social media to focus on the potential hazards around them. “It feels like the set of a reality show,” he says. “They aren’t paying attention to personal safety.”

Dorn, whose travels have taken him to troubled regions all over the world, says that all researchers should go online to check for official travel notices — such as the foreign travel advice at go.nature.com/zioaff or alerts and warnings from the US Department of State — before heading to potentially hazardous areas.

Travel to specific countries is rarely banned outright, but there are often many things to keep in mind. For example, the state department currently advises drivers in Mexico to reduce the risk of carjacking by driving between cities only during daylight hours and using toll roads whenever possible.

A camp that is too remote for Wi-Fi or mobile-phone reception should at least have a satellite phone, Dorn says. Even then, however, the necessary evacuation might not be possible if the PI doesn’t have medical rescue insurance. Dorn says that insurance policies

“I would rather drive someone round 500 times than investigate a rape.”



Michael Danti's team at the tomb complex of Gund-i Topzawa (400 BC) in northeastern Iraqi Kurdistan.

are worth double-checking: some cover rescue landings and take-offs only if they are from paved airstrips, which are not always an option out in the field. “You have to ask, ‘If something happens at this particular site, will somebody come to get me?’”

Some PIs need to rethink the casual, anything-goes culture that seems commonplace at field sites, says Katie Hinde, a human evolutionary biologist at Harvard University in Cambridge, Massachusetts. “It’s like an adventure, a vacation or a camping trip,” she says. “There’s a relaxation of professionalism.” She says that in conversations with other scientists and during interviews with people participating in a study she co-authored called Survey of Academic Field Experiences (SAFE) report (K. B. H. Clancy *et al.* *PLoS ONE* 9, e102172; 2014), she has heard many variations on the same story: researchers who were generally respectful in the lab seemed transformed in the field, leading to inappropriate behaviour and unwanted physical contact. And in fact, one-fifth of the 516 women and 150 men who participated in the study said that they had been sexually assaulted during a field trip. Several said that they had experienced multiple assaults.

Protecting scientists from one another can be a tricky and delicate task, says Hinde. As a first step, she says, PIs should remind all team members (and themselves) that the university’s code-of-conduct rules apply to the field as well as the lab. Yet, many seem to be silent on such issues, especially in the field. Fewer than two-fifths of the people who responded to the SAFE survey recalled ever having been briefed about rules of conduct. “PIs shouldn’t feel obligated to follow those rules just because it’s the law,” she says. “They should do it because they care about their colleagues.”

Hinde believes that safety considerations regarding colleagues as well as outsiders should be part of any grant application for field work. It would be an official acknowledgement of a problem that should not be ignored. “This

hadn’t been on people’s radars because people didn’t talk about it,” she says. “Now a lot of people are working very hard to make their field sites safe.”

USE INITIATIVE

Dorn, a former police lieutenant at Mercer University in Macon, Georgia, says that some researchers seem to court trouble by working alone and keeping late hours. He says that a police officer on his team once pulled a gun on Mercer’s president after mistaking him for an intruder in the middle of the night. After that, the president vowed to let security know when he would be working late. Dorn also recommends that researchers who need to work late at night check with campus security for advice and support, or for a ride, if it does not seem safe to walk into a car park or across a vast, dark campus. “I would rather drive someone around 500 times than investigate a rape,” he says.

Dorn says that one strategy researchers can use to help improve their safety is visualization. The idea is to calmly visualize every potentially dangerous situations and decide ahead of time how to react. That way, “you’re preloading your brain with information,” he says. Law-enforcement officers, soldiers and athletes have been using the technique for decades to map out strategies for various what-if scenarios. With a plan in place for each one, he says, they can proceed with confidence and alacrity.

Dorn and his colleagues discuss visualization and other safety strategies at the website safehavensinternational.org/staying-alive.

The potential for danger has not stopped Danti from pursuing his passion. “I love Syria, Iran and Iraq,” he says. “I love the people and the culture.”

Even more than that, however, he loves bringing his team home safely. And once he is back in his office, he can start thinking about everything that can go wrong next time. ■

Chris Woolston is a freelance writer in Billings, Montana.

MICHAEL DANTI

TURNING POINT

Stacey Gabriel

MARIA NEMCHUK
US genomicist Stacey Gabriel was named 'hottest researcher' on Thomson Reuters' World's Most Influential Scientific Minds 2014 for publishing 23 of the most highly cited papers in 2013, the most recorded for the year. She directs the genomics platform at the Broad Institute in Cambridge, Massachusetts.

For how many papers were you lead author?

None — which speaks to the highly collaborative nature of the work we do here.

Was it difficult to publish this many highly cited papers?

With so many good ideas springing up for projects, the challenge is keeping up with what technical advances need to come next. One of the Broad's founding principles is to make possible the types of project that are not feasible in individual research labs. We built sequencing and microarray capabilities that enable very large projects. For example, most of the papers I am working on right now are about identifying new regions of genes that may be implicated in a specific disease. The research bottleneck happens at the next step, which is the follow-on research that investigates which genome variations result in cancer.

Do you have a recent favourite?

Nature accepted one in late 2013 for publication in January 2014 (M. S. Lawrence *et al. Nature* **505**, 495–501; 2014) in which we looked at many tumour types to search for undiscovered cancer-causing genes. We found a few that we had not appreciated before, including some involved in cell death, genome stability and RNA processing. It brought together an enormous amount of work.

What is it like to be the most influential scientific mind of 2013?

Honestly, I don't keep track of it. I was amused that I had three more publications than my boss, Eric Lander — that did make me chuckle. But my career has revolved around building projects, not keeping track of my publication record.

How did you come to pursue genetics?

I was working as a phlebotomist at the University of Pittsburgh in Pennsylvania when I started studying diseases in Mennonites, a religious denomination that is considered a genetically isolated population. As a graduate student, I was drawing blood and taking family histories from participants. I learned how to isolate DNA from blood and



do mapping studies, which prompted my interest in this field.

Did the Human Genome Project (HGP) influence your career path?

Yes. I was completing graduate school in 1998, when genomics and human genetics were taking off. There were opportunities at many institutions to study DNA variation, which interested me. I met Eric — who has been my boss for 17 years — before the Broad Institute existed. While I was a graduate student, he gave a lecture that completely hooked me. Because my adviser knew him, I was able to meet him a few years later when I was looking for a position. He was developing some of the first microarrays to study human polymorphisms at a large scale, which was incredibly appealing. I've been studying DNA variation ever since.

What do you see ahead for 2015?

We're really at the tip of an iceberg when it comes to surveying the genomic landscape of many cancers. We're seeing how the application of technological advances to extensive sample collections sets the stage for discoveries. This rapid pace of discovery won't be a blip, given how fast technology is advancing.

Which paper do you consider a turning point?

In 2002, early in my career at Broad, we had a *Science* paper about haplotype blocks — sets of inherited DNA variation — in the human genome (S. B. Gabriel *et al. Science* **296**, 2225–2229; 2002). I was lead author, and it turned out to be a high-profile paper that helped me to become known to the broader community and write my own grants. ■

INTERVIEW BY VIRGINIA GEWIN

COMPLIANCE

Research regulations

The US National Academies is examining research governance for its impact on universities and researchers. A committee of experts on higher education, science and policy will work with university investigators to determine the labour and costs needed to comply with reporting requirements and other regulations, and to identify areas in which the added workload outweighs the benefits of compliance. Project director Anne-Marie Mazza says that her committee will gather input from regulators and investigators to understand the initial reasons for creating regulations, and to evaluate ongoing needs and implementation in light of current research practices. The study's findings are expected to be published next year.

MOTHERHOOD

Biases in US academia

Organizational biases against motherhood exist in academia, find US researchers. Kirsten Isgro at the State University of New York Plattsburgh and Mari Castañeda at the University of Massachusetts Amherst found that US universities and colleges conflate sabbatical leave and maternity leave, or expect female faculty members to time their child-raising years around tenure decisions (K. Isgro and M. Castañeda *Womens Stud. Int. Forum* <http://doi.org/z3j>; 2015). They collected accounts from more than 300 women in academic positions. Isgro says that there is no single ideal path for a woman to mesh her life as a parent with her work.

CHARITABLE DONATIONS

University gifts grow

US universities received more revenue from philanthropy and investments in 2014. A survey from the Council for Aid to Education in New York found that charitable contributions hit a record high of US\$37.5 billion, up 11% from the previous year, and the biggest jump since 2000, when donations rose by 14%. And a survey by the National Association of College and University Business Officers found that endowments netted an average return of 16%, up from 12% in 2013. The rapid growth is making up for difficult years after the financial downturn in 2008, says economist Richard Freeman, who directs the National Bureau of Economic Research in Cambridge, Massachusetts.

GOOD FOR SOMETHING

Complete control.

BY DEBORAH WALKER

Raoul reached in his pocket and touched the cold copper of the Loonie coin. It would be safer in his wallet, or safer still back on Earth with the rest of his collection. But he needed to touch it. The Loonie bought him luck. He always carried it when he was acquiring. The Loonie had been his first coin. It was right he should have it when he obtained his last.

A woman smiled at him. She was probably a distraction for a pickpocket. When he'd arrived at High Jova, they'd warned him that the orbital teemed with thieves. Why would else would an attractive woman smile at him? Then he remembered he was wearing nano-skin. He looked 60 years younger. Young, handsome, just like he'd looked when he'd first met Sven.

But he walked on scowling, hand tight around the Loonie.

The Loonie. A hard-times coin, issued when the First Lunar Bank went bust in 2101. It was the unofficial currency of the lunar depression. Electronic credit was fine, convenient. But there would always be times when people needed money; needed hard, cold cash.

The Loonie was his first coin, acquired by accident half a century ago and it had sparked a desire in Raoul, the start of a life time's obsession with collecting coins from every off-world habitation. He thought of his home on Earth, where his collection rotated elegantly in the display field. Coins were history in a way that electronic credit could never be. Not only social-economic history, but the personal history of what Raoul had done to acquire a particular coin. Sven didn't understand that.

Raoul left the bustle of the market sector and turned into the quiet corridor of the living quarters. In his wallet was the card-key that the Slider had assured him would get him into Ben Dell's quarters. She'd sneered, but she'd taken his money. It would get him in. A Slider's word was as cold and certain as vacuum.

Dell owned the largest coin

collection in private hands. But unlike Raoul, Dell bought coins indiscriminately. It was unfortunate then, that the elusive Titan Good For coin had found itself into Dell's hands.

The Titan Good For had been issued to workers building the first Titan orbital in 2128, before the disastrous core meltdown. A Titan Good For coin had been exchanged for one meal in the workers' canteen. It had been thought for many centuries that these coins had been lost in the destruction of the station. Raoul had a fine Titan Two coin, those were two a penny. But when he heard that Dell had retrieved a Titan Good For from a derelict escape pod, he knew that he had to have it. With a Titan Good For, Raoul would have a complete set of all coins issued in off-world habitations during the twenty-second century.

Which was why Raoul was on High Jova, wearing nano-skin, carrying Dell's card-key in his wallet. He'd tried to make Sven understand.

"If I complete the collection, I'll have done something with my life."

"We've been married for fifty years. We have two children. We have seven grand kids. You've done something."

"I know. But this is something special."

"It's a hobby."

"It's not just a hobby."

"Don't go to High Jova, Raoul. Don't throw your life away."

Dell had been so unreasonable. Raoul had approached him, through a number of intermediaries, offering a fair price, then a ridiculous price. Then he'd liquidated all his assets and offered an extortionate, exorbitant price, even though Sven had begged him not

to. But Dell had rejected all offers. In the end, Raoul had made a personal appeal to Señor Dell, collector to collector. But Dell had

refused out of spite.

Dell wasn't interested in twenty-second-century off-world coins. He didn't even have the semi-

rare Mars Mark. He just wanted to stop Raoul achieving his dreams.

Sven told him that he was crazy.

Raoul had tried to explain. He really shouldn't have told Sven what he was planning to do.

Sven had given him an ultimatum. "Don't do this. Let it go. Otherwise..."

Otherwise?

Sven had left him. After 60 years. He didn't understand.

Raoul was very close to Dell's quarter, his heart beating wildly. He was about to risk everything. If Sven could feel what he felt, maybe he'd understand. Essentially Raoul was doing this for Sven. With the Titan Good For, his collection would be complete. Complete. Then Raoul would be able to fix everything, remind Sven of the man he'd been 60 years ago, of the love they'd had, the life time they'd shared, the future they could have together. But he had to have the Titan Good For. History had to be complete. It all had to make sense.

Raoul reached the door. He took the card-key from the wallet. The Titan Good For would be inside, and maybe Señor Dell would be inside, too. Raoul carefully drew the gun out of his bag. The Sliders had supplied that, too.

Sven didn't understand. Coins were history, and history had to be complete. It all had to make sense. One last time, Raoul touched the Loonie in his pocket. Hard times. A collection had to be complete to make sense. A life had to be complete. A life had to be good for something. ■

Deborah agrees with Ralph Waldo Emerson: money often costs too much.

ILLUSTRATION BY JACEY

